Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities

Angelina McMillan-Major

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Emily M. Bender, Chair
Batya Friedman
Gina-Anne Levow
Shane Steinert-Threlkeld

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

Language Dataset Documentation Design: Learning from Deaf and Indigenous Communities

Angelina McMillan-Major

Chair of the Supervisory Committee:

Emily M. Bender
Department of Linguistics

This dissertation investigates how engaging with stakeholder groups, namely natural language processing (NLP) practitioners and language communities, can contribute to the development of documentation toolkits that are more responsive to the needs of these groups. The development process follows value sensitive design in conducting a series of investigations to learn what are the needs of these groups and how iterative improvements to technology can help address those needs. Building from the data statements for NLP Version 1 schema proposed in Bender and Friedman (2018), Dr. Emily M. Bender, Dr. Batya Friedman, and I conduct an empirical investigation and a technical investigation to develop the data statements Version 2 schema by engaging with natural language processing professionals. To learn about the needs of indigenous and deaf communities with respect to collaborating with researchers, in a retrospective technical investigation I analyze ethical guidelines and licenses for the values frequently expressed in these communities' stated expectations for research collaborations. I then conduct a technical investigation to meld the data statements Version 2 schema, aspects of datasheets for datasets (Gebru et al., 2021), and the results of the retrospective technical investigation into a single toolkit. Rather than documenting existing datasets, the Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources (C3DAR) toolkit is designed to facilitate collaborative partnerships between communities and researchers working to develop language datasets. I conclude with possible future investigations, focusing on community researchers as key stakeholders, and considerations for uptake.

Contents

List of Figures

List of Tables

List of Acronyms

AI	artificial intelligence
ASL	American Sign Language
CS	computer science
IDS	indigenous data sovereignty
ISE	International Society of Ethnobiology
ML	machine learning
NLP	natural language processing
SLP	signed language processing
UNCRPD	United Nations Convention on the Rights of Persons with Disabilities
UNDRIP	United Nations Declaration on the Rights of Indigenous Peoples

Acknowledgements

First I'd like to express my gratitude and appreciation for my committee, Emily M. Bender, Batya Friedman, Gina-Anne Levow, and Shane Steinert-Threlkeld, whose lessons over the years and suggestions for this work greatly improved my skills in thinking and writing about language and technology. I especially want to thank Emily for her guidance and patience with me over the years as I meandered my way through graduate school and took too many language courses. Thank you to Emily and Batya for inviting me to join you on this journey with data statements. It has been a privilege to learn from you both.

Thank you to Anna Lauren Hoffman for kindly taking on the role of Graduate School Representative, and for a class that first introduced me to critical data studies.

My sincerest thanks to the participants of the 2020 LREC workshop and to the experts I spoke with about C3DAR for sharing their time and experiences with me: Benjamin Frey, Julie A. Hochgesang, Rose Stamp, Santiago Esteban, and Merve Ünlü Menevşe.

Thank you to everyone from the Linguistics department who helped to create a community that I will very much miss: Kristen Howell, Olga Zamaraeva, Courtney Mansfield, Amandalynne Paullada, Marina Shah, Naomi Tachikawa Shapiro, Sara Ng, Rob Squizzero, Mike Haegar, Eli Miller, Ajda Gokcen, Alex Spivey, Mike Furr, Joyce Parvi, and many others.

Thank you to Ryne MacBride for coffee shop writing sessions.

Thank you to everyone at the Language Learning Center who first encouraged me to consider applying to the PhD program, especially Paul Aoki, Russell Hugo, Michele Anciaux Aoki, Sherri Huber, and Larry Lesage.

Thank you to my family for their love and support, especially Kade Jones, Brian Jones, and Kylie McMillan.

Thank you to Tatyana Sasynuik, Taylor Bettine, Hanne Puntervold, John Richards, Brenna Moore, Erik Sorvik, Cody Taylor, Julia Taylor, Helen Jones, and especially to Zach Eley for always cheering me on and reminding me there is in fact life outside of graduate school. I would not have been able to complete this program without you all. Thank you, Tatyana, for always being willing to listen, helping me communicate my ideas better, and giving me space to vent when I needed it.

Last but not least, thank you to my cat Papaya, for reminding me to take breaks.

I am grateful to have been supported financially by the UW Language Learning Center, the UW Department of Linguistics Excellence in Linguistic Research Graduate Award, the UW Tech Policy Lab, and the Frances and Howard Nostrand Endowment.

DEDICATION

For my grandmother, Yvonne Lorraine de Jong

Chapter 1

Introduction

> *We are aware that researchers are knowledge brokers, people who have the power to construct legitimating arguments for or against ideas, theories or practices. They are collectors of information and producers of meaning, which can be used for or against Indigenous interests.*
> - Nuu-chah-nulth Tribal Council Research Ethics Committee (2008)

Research is an inherently political activity. The questions that are asked, the methods that are used, and even who asks the question all impact what observations may be made and how they are interpreted. The illusion of the objectivity of science is not new, but the advancement of Big Data has reignited "the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy" (boyd and Crawford, 2012, pg. 663). Governments in particular have hailed the phenomenon of Big Data to support more responsive public policy. The United States' President's Council of Advisors on Science and Technology lauded Big Data and the potentials of digital democracy "to improve public discourse, increase dialogue between citizens and government, make government more open and transparent, improve the operation of government, and bring the benefits of technology to everyone" (President's Council of Advisors on Science and Technology, 2010, pg. 33). For some, the increasing capacity of governments and companies to collect vast amounts of data on individuals has sparked new concerns about privacy. For others, namely minoritized communities, government overreach and subjugation through data continues longstanding struggles for self-determination (Kukutai and Walter, 2015; Carroll et al., 2019).

After centuries of hegemonic assimilation practices, minoritized communities are increasingly leveraging collective power and technical skill for gathering and processing their own data so that communities may make informed decisions regarding their own futures and advocate for themselves against technologies that are not built to address their needs.[1] Technologies that are put forth as the ultimate solution to embedded social problems instead discriminate against minoritized communities due to incomplete understandings of the context of the problems on the part of most tech developers and due to data and algorithmic designs that promote stereotypical representations of these communities or ignore them entirely. Green (2019) points out that computer scientists lack methodologies for measuring the benefits of technologies designed for social good, using criminal justice reform as an example frequently targeted by technologists. In the field of computer vision, Buolamwini and Gebru (2018) show that a facial recognition system achieved higher accuracies for individuals with lighter skin based on the Fitzpatrick skin classification system and for men, while it performed worst for women with darker skin tones. For natural language processing (NLP), Bender, Gebru et al. (2021b) highlight the harms of large language models, including the resulting environmental impacts that training language models has upon communities whose languages are generally not supported in language technology. Language technology in particular presents risks originating from the general relationship between language and identity (Eckert and Rickford, 2001) as it affects model performance for minoritized groups (e.g. Wassink et al., 2022), the propensity for humans to attribute meaning to fluent language they encounter even if they know it is machine-generated (Bender and Koller, 2020), and the dearth of NLP research for the majority of the world's languages (Joshi et al., 2020).

Where the governments, academia, and tech companies have failed to meet the needs of these communities, community researchers and community-researcher collaborations understand that "the data that we handle are human fates" and that automated, flattening data processes do not treat individuals with the dignity they deserve (Raji, 2020, pg. 2). These collaborations have brought attention to the ways that communities have been failed and have bridged data gaps in the ways that Big Data continues to shift economic power away from Black people (Milner and Traub, 2021), in the air pollution disproportionately affecting low-income communities (Bayram and Hersher, 2023), and in the systematic barriers deaf people face in

[1] See e.g., Data for Black Lives (`https://d4bl.org/`), Collaboratory for Indigneous Data Governance (`https://indigenousdatalab.org/projects-working/`).

succeeding in higher education.[2] In indigenous communities around the world, this effort towards the governance of community data has been termed indigenous data sovereignty (IDS) (Walter et al., 2021). To work towards this goal, Holton et al. (2022) identify two needs: 1) indigenous-led research and data management planning and 2) infrastructure and resources.

Documentation has emerged as one of many tools for managing data and mitigating the harms of technology built off of data. The proposals vary in terms of scope and the targeted aspects of the technological development process to be documented. FactSheets provide information about the purpose, performance, safety, security, and provenance of AI services (Hind et al., 2018; Arnold et al., 2019). Model cards report technical details of model algorithms and encourage intersectional benchmark evaluations (Mitchell et al., 2019). Two documentation formats for dataset documentation make use of the ingredient and nutrition label metaphor to visually present at-a-glance information about the contents and recommended uses of datasets, Nutrition Labels for Data and Models (Stoyanovich and Howe, 2019) and the Dataset Nutrition Label (Holland et al., 2018; Chmielinski et al., 2022). Datasheets for datasets (Gebru et al., 2021) and data statements for NLP (Bender and Friedman, 2018) both produce long-form prose documentation specifically for datasets used in machine learning (ML) systems, independent of the models and systems that rely on datasets, with the latter format designed with specific questions for language data types. In each case, the primary goal of the documentation is to support accountability and transparency measures by enabling parties outside of the development team to find answers relating to the provenance and design of the system or dataset.

Documentation toolkits and proposals have so far primarily been designed to support documentation authors and readers who are academics and industry professionals in technical and technology-related fields. This thesis will address the following research questions: How can we develop tools and practices for dataset documentation that are responsive to the needs of more varied stakeholders groups? How can documentation support collaborative language technology work between researchers and language communities? In this work, I propose a toolkit, called the C3DAR (Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources) toolkit, pronounced "cedar" like the tree. The C3DAR toolkit consists of a specification and a set of practices for planning the creation of language datasets with communities grounded in the collaborative and community-led research frameworks. The C3DAR toolkit and template is available

[2]National Deaf Center on Postsecondary Outcomes (`https://nationaldeafcenter.org/news-items/get-strategies-deaf-led-system-change-new-resource/`)

for download at https://digital.lib.washington.edu/researchworks/handle/1773/50585.

I begin by providing background on the bodies of literature that I draw on for this work (Chapter 2). In §2.1 and §2.2, I summarize the hegemonic research practices that indigenous communities and signed language communities, particularly in the United States, have endured, and how the campaigns for indigenous and deaf linguistic and cultural rights have changed ethical practices in research fields concerned with language and contributed their own lineages of research. Language technology has increasingly been developed for indigenous and signed languages, however the fields of research that focus on language technology are in the midst of shifting their own ethical research practices. I outline the ethical concerns and frameworks in NLP and ML in §2.3 and explain how documentation has emerged as a tool for mitigating the risks of technology, and particularly language technology. In §2.4 I give a brief overview of value sensitive design as the methodological foundation of this work and frame the following work using the value sensitive design terminology of conceptual, technical, and empirical investigations.

Chapter 3 describes how my co-authors, Dr. Emily M. Bender and Dr. Batya Friedman, and I improved upon one of the existing documentation toolkits, data statements for NLP (Bender and Friedman, 2018). Using value sensitive design methods, we first conducted an empirical investigation; in this investigation we held a workshop with language technology developers and collected feedback on their experiences with writing data statements. Then, in a technical investigation, we improved the schema using the workshop feedback and compared the interim data statements schema to datasheets for datasets (Gebru et al., 2021). These investigations produce the data statements Version 2 schema presented in McMillan-Major et al. (2023) and a guide for writing data statements (Bender et al., 2021a).

Chapter 4 presents my methodology of two investigations for incorporating the values of indigenous communities and signed language communities into the development of C3DAR, a retrospective technical investigation and a technical investigation. Chapter 5 reports on the results of the first of these. I analyze ethical guidelines and licenses published by language communities from around the world and discuss the values that are frequently expressed across these documents. In Chapter 6 I describe the results of my next investigation, a technical investigation, in which I again leverage datasheets for datasets as well as the results of the retrospective technical investigation to develop the first iteration of C3DAR from the data statements

Version 2 schema. The limitations of this work and the tensions inherent in trying to address the varied needs of so many stakeholder groups are discussed in Chapter 7. To learn more about the tensions and experiences of specific communities, Chapter 8 proposes a future empirical investigation and reports on initial conversations with community member researchers. I conclude in Chapter 9 with possible use cases of documentation toolkits like data statements and C3DAR as well as future work and open questions for collaborative dataset documentation.

Chapter 2

Literature Review

Political activism for the cultural and linguistic rights of communities and the efforts of community member researchers and allied researchers have shifted the standards of ethical research toward community-led research and more collaborative research frameworks. At the same time, communities and allied researchers have called for language technology to support access to and access in indigenous and signed languages (e.g., Adda et al., 2019). That being said, technical innovations are best positioned to address the needs of the intended audience when they are developed with an understanding of the social and historical contexts of those audiences. In this chapter, I provide brief histories of the complex relationships indigenous communities (§2.1) and signed language communities (§2.2) have with academic and government institutions, particularly within the United States. In §2.3, I present the ethical considerations that researchers in technical fields bring to the conversation on language technology. Recent language technologies have been fraught with improper data management and compensation practices, deployed without consideration for the possible negative impacts of the technology's use, and often simply framed as the answer to a problem that has no easy solutions. Documentation has emerged as a tool for mitigating the some of these harms and supporting accountable data practices. All the calls for interdisciplinary design with communities beg the question of how we design documentation or any other technology with an eye towards positioning communities as leaders and designers. I introduce value sensitive design (§2.4) as the methodology I employ to explore how documentation can support language community and research collaborations to create data designed to support communities' self-defined language goals.

Researcher Stance I am an American computational linguist. I am of Chippewa-Cree and white European descent and my connection with my indigenous heritage is grounded in my connection with my maternal grandmother, a Chippewa-Cree tribal member. I was born in Honolulu, Hawai'i but have lived most of my life in the ancestral homelands of the Coast Salish peoples and the Snoqualmie tribe, in what is now known as Washington State. I am hearing and have had some exposure to American Sign Language (ASL) and deaf culture from my undergraduate courses. Because of my position as an academic in the context of the United States, I primarily focus on the deaf and indigenous communities in the United States, with the recognition that these communities are not uniform or homogeneous across the nation, nor are they necessarily representative of deaf and indigenous communities around the world. I do not presume to represent any of these communities, indigenous or deaf, but rather hope that my work can help make the work that has been done and is being done by these communities more visible.

I draw from many interrelated technical fields. As a computational linguist, I use technology and computational methods to investigate questions about language and linguistic theory. Computational linguistics overlaps and is sometimes seen as equivalent to natural language processing (NLP). NLP research has grown to encompass more research on technical applications of language and focuses less on linguistic theory. NLP borrows methods and algorithms from machine learning (ML). ML includes algorithmic and engineering research for replicating general patterns, which may be designed explicitly in supervised ML or surface implicitly in unsupervised ML, from training data. Technology built by technical practitioners from ML, NLP, computer vision and other technical fields are sometimes referred to as artificial intelligence (AI) when the technology addresses complex problems using algorithmic pattern-matching and decision-making processes. AI ethics has emerged as a field dedicated to the critical analysis of AI technologies and their societal implications. My work pertains to practices that position technologists to support community values in creating language technologies, thus falling under NLP ethics, which is closely aligned with the broader study of AI ethics.

Terminological note Within deaf studies, capitalization of *Deaf* has been used to reference the sociocultural aspects of the deaf experience in contrast with *deaf*, used to reference the biological aspects (Woodward, 1975; Woodward and Horejes, 2016). While many deaf scholars continue to use the notation, others have presented arguments against its use. Kusters et al. (2017a), quoting Woodward and Horejes (2016)'s

recount of the coinage, note that the usage of the d/Deaf distinction has come to represent more ideological notions of who is or is not culturally Deaf. Kusters et al. (2017a) do not use the d/Deaf distinction, calling it "paternalistic, obscuring, and imposing" for academics to label someone else, even someone who is deaf and uses a signed language, as culturally deaf when they may or may not use that label for themselves (pg. 14). Bragg et al. (2021) point out that the d/Deaf distinction creates a false sense of uniformity across signed language cultures, and for this reason, they do not capitalize usages of *deaf*. Following Kusters et al. (2017a) and Bragg et al. (2021), I use lower case *deaf peoples* or *deaf communities* in all instances to refer to various groups of people who experience biological or corporeal deafness, rather than in reference to the sociocultural and medical models of deafness (see §2.2.1), except when quoting prior work that does use the D/deaf distinction. This includes hard of hearing and deafblind experiences of deafness.

The use of *Indigenous* versus *indigenous* varies geographically (see Peters and Mika, 2017, for a discussion of *indigenous* as opposed to similar terms). I see upper case *Indigenous* used most often in North America, where indigeneity is racialized (discussed further in §2.1.3) and used in contrast with other racial categories like Black, Asian, and Latina (Oliver, 2017). The term *indigenous* is not universally accepted (Gagné, 2015) and is not defined in the United Nations Declaration on the Rights of Indigenous Peoples because a universal definition is not appropriate (Asia Pacific Forum of National Human Rights Institutions and Office of the High Commissioner for Human Rights, 2013). Instead, each indigenous group has the right to define themselves (Office of the High Commissioner for Human Rights, 2007). Following Smith (2012), I use *indigenous peoples* to refer to the marginalized peoples who have been subject to colonialism and imperialism, where the members of each group of people identify with a shared linguistic and cultural background. Since I refer to indigenous communities around the world, and in keeping with the arguments against essentialism from the previous paragraph, I use lower case *indigenous peoples* to avoid imposing implications for identity where they may not be appropriate and to refer to the multitude of unique groups who identify as indigenous.

I use *community/communities* throughout this dissertation to refer to groups with shared knowledge. For signed language and indigenous communities, the shared knowledge is often cultural knowledge and linguistic knowledge of specific languages. Signed language communities overlap with deaf communities, in that there are many deaf signed language users, but signed language communities also include hearing

people who sign such as the relatives of deaf individuals, interpreters, and second language learners. See De Meulder et al. (2019a) for in depth discussion of signed language communities. Peters and Mika (2017) note that indigenous peoples tend to prefer their own collective names though these may be contested. To refer to both of these groups collectively, I use the term *minoritized language communities* as communities whose languages have been subjected to policies resulting in fewer speakers of the community language. For academic communities such as linguistics and natural language processing, the shared knowledge is instead technical and theoretical. In each instance, *community* is a convenient term that does not do justice to the diverse, nebulous, and ever changing nature of any group of people. Throughout this work, I am guided by Hochgesang and Palfreyman (2022)'s cautionary advice: "notions of 'involving the community' must be qualified by limitations on how far any one signer (or groups of signers) can represent or stand for the totality of people who use a signed language variety" (pg. 159).

2.1 Indigenous Data Governance and Ethics in the Context of Language Technology

In this section I provide a brief history the complex relationship indigenous peoples have with research, particularly indigenous peoples in the United States. This history begins the hegemonic research that was conducted on indigenous communities that objectified them and subjected them to colonial powers (§2.1.1). In the United States, the politics of genocide and assimilation have abated, however federal efforts toward the revitalization of indigenous languages and cultural practices have produced mixed results (§2.1.2). Language research has also contributed to exploitative practices against indigenous communities; linguists who are themselves indigenous and linguists who have long term working relationships with indigenous communities have made clear the harms of these practices and the benefits of more collaborative models of research, changing perceptions in the field about the positions of researchers with respect to community members (§2.1.3). Indigenous people are increasingly leading research and policy development following their own methods and values and in support of their own community needs using frameworks such as Indigenous Data Sovereignty (IDS) (§2.1.4). In §2.1.5, I consider the implications of the colonial history of language research and the current topics voiced by indigenous community members on the paths forward

for indigenous language technology development and sovereignty.

2.1.1 Hegemonic Research in Indigenous Communities

Research in its various historical and modern forms has long been complicit in enacting and validating colonialism against indigenous communities (Smith, 2012). The first to document indigenous languages were often missionaries from imperialistic nations who aimed to develop orthographies for translating the Bible and convert indigenous people to Christianity (Errington, 2008). Religious, political, and economic factors motivated invading European forces' violent dispossession of land, culture, and language from the indigenous communities, but scholars rationalized this violence by putting forward dehumanizing theories about non-European peoples, including indigenous peoples. Early scholarly work on linguistic and cultural relativism that theorized hierarchies of languages in which European languages were somehow superior "contributed to the sense of otherness of colonial subjects whose humanity could be counted as incommensurable with that of their masters" because qualities of languages were assumed to reflect broad qualities of the people who spoke those languages (Errington, 2008, pg. 51). These disparaging notions of indigenous languages and cultures also contributed to the genocidal assimilation policies of colonial nations like the United States, Canada, and Australia. Such policies attempted to eradicate indigenous peoples by forcibly placing indigenous children in boarding schools and residential schools in order to strip them of their cultures and languages and thereby also erase their indigenous identities (Bone et al., 2022). These policies were grounded in early philosophies connecting language, culture, and identity whose later iterations underpin much of modern linguistic research.

Western scientific endeavors continued to dehumanize indigenous people through practices that treated the researched as objects. Underpinning these interactions is *scientific colonialism*, the belief that Western scientists have the right to any data source and information and right to export that data from marginalized communities for the greater benefit of humanity (Nobles, 1976; Cram, 2006; Chilisa, 2020), despite the harms that may be incurred by those marginalized communities and the lack of immediate benefits to them. Organizations in biomedical fields have conducted research without informed consent (National Research Council et al., 1996) and contributed to stereotypes of indigenous people (Hyett et al., 2019), resulting in reduced quality of care for indigenous people. Indigenous communities in Africa have long been fighting

for the rights to their indigenous knowledge of plants such as the rooibos plant (National Khoisan Council & Cederberg Belt Indigenous Farmers Representatives, 2019), used as a tea and for other medicinal purposes, and the hoodia cactus plant. This cactus was long used by the San to stave off hunger, but the components of the plant that produce this effect were isolated by a UK-based pharmaceutical company and sold as a diet pill (Commey, 2003). This pattern of indigenous knowledge of plants patented and marketed by companies based in Western countries is now known as biopiracy and is also rooted in colonial practices.

In linguistics, extractive methods for collecting data focus on decontextualized language rather than the people who embody the language. This framing allows linguists to claim ownership of the language through copyright of publications (Davis, 2017; Gaby and Woods, 2020). Such efforts have increased following calls to halt the process of language extinction and save endangered languages (Hale et al., 1992; Krauss, 1992). Despite acknowledging the compounding social, economic, and historical factors contributing to language loss (such as the religious motivations of some linguists contributing to language documentation as described in Erard, 2005), the reduced number of speakers of indigenous languages has been framed by some researchers as an intentional choice by community members (e.g., Ladefoged, 1992). Even linguists who advocate for more support for communities working to revitalize their own languages still sometimes use argumentation that alienates the language community; these include claims of universal ownership of languages, hyperbolic phrases about the value of languages, and the use of statistics which requires imposing essentializing categories on languages and speakers (Hill, 2002). By starting from these motivations, the focus of linguists' efforts have at times shifted to saving the language as data and away from supporting communities in their own ongoing efforts to reclaim their languages by teaching it to community speakers of all levels (Leonard, 2017; Bird, 2020). The idea of "saving languages" has been criticized as normative for applying an assumed universal solution to a highly context-sensitive problem (Dwyer, 2006).

2.1.2 US Government Support for Indigenous Languages

Governments and researchers in the United States have made many changes in how they interact with indigenous communities. In 1990, the United States government passed the Native American Languages Act which establishes that the federal government will "preserve, protect, and promote the rights and freedom of Native Americans to use, practice, and develop Native American languages" (Native American Languages

Act, 1990). It also encourages the use of local languages in both primary and secondary education; funding for primary education in indigenous languages and language preservation programs was later provided by the Esther Martinez Native American Languages Preservation Act (2006). Five years later, the Linguistic Society of America called for renewed government support for indigenous languages, citing further need for resources to achieve the goals the prior acts had established (Linguistic Society of America, 2011). Provisions for indigenous children's education with an emphasis on culture were included in the Every Student Succeeds Act (2015) and more funding was provided with the reauthorization of the Esther Martinez Native American Languages Preservation Act in 2019.

Despite these efforts, surveys of American Indian and Alaska Native students find that more support for indigenous language education is needed, especially for students in public schools with less than 25% of their student population identifying as American Indian and Alaska Native. In 2015, the National Indian Education Study found that 47% of grade 4 students and 52% of grade 8 students reported never having exposure to any indigenous language (Rampey et al., 2019). When students were asked about their knowledge of their own tribes and groups, 13% of grade 8 students responded that they knew "nothing" about their tribal or group's history, traditions and culture, or current issues. This figure increased in a following study in 2019, where 17% of all American Indian and Alaska Native students in grade 4 and 18% of all students in grade 8 respond that they knew "nothing" about their tribal or group's history, traditions, or arts and crafts (Rampey et al., 2021).

Most recently, the federal government has approved new funding mechanisms for improving and expanding indigenous language pedagogy and collecting indigenous language statistics through the Native American Language Resource Center Act (2023) and the Durbin Feeling Native American Languages Act (2023). The Native American Language Resource Center Act provides grants for institutions to establish, strengthen, and operate a resource center and to staff that center with indigenous language experts. The act states that a center "must serve as a resource to improve the capacity to teach and learn Native American languages, further Native American language use and acquisition, and support the revitalization and reclamation of Native American languages." The Durbin Feeling Native American Languages Act authorizes the federal government to collect data and report on the number of indigenous languages in the United States which are currently spoken, the estimated number of speakers, and other related statistics. The stated intent

of this data is to improve the efficiency of the federal government in supporting indigenous communities' language reclamation efforts. It remains to be seen the degree to which these pieces of legislation support indigenous self-determined goals, as evidenced by the proportion of *indigenous* experts of indigenous language and pedagogy who receive funding and the amount of control indigenous communities have over the data collection and management processes for these surveys.

2.1.3 Linguistics and Indigenous Communities: Frameworks for Engagement

Since the extractive practices described in §2.1.1, linguists have reflected on the relationships that the field has had with the communities they have worked with. Cameron et al. (1992) characterize social scientists' research based on their attitudes towards research subjects over the years using three different models: *ethical research*, *advocacy research*, and *empowerment research*. They further describe these models as "research *on*", "research *on and for*", and "research *on, for, and with*" communities, respectively (Cameron et al., 1992, pg. 22).

Rice (2006) illustrates each of these models with quotes from various linguists over the years whose work in language documentation provide examples of Cameron et al.'s (1992) models. The descriptions and terms used to address the people from the community who assist the linguist exemplify the changing relationships between linguistics and communities over the years. In ethical research, community members are referred to as "informants" and the primary ethical concern is how the community member is compensated for their work. For advocacy research, linguists additionally have a responsibility to support the community in ways that align with the linguists' research goals. Empowerment research goes a step further and includes community members in the research, "taking into account the knowledge that the speakers bring and their goals and aspirations in the work" (Rice, 2006, pg. 132). Rice (2006) exemplifies the shifts in linguist-community relations in later frameworks using the changes in accepted terminology for referring to community members; the previously used term "informant" came to be seen as pejorative to community members, leading instead to the use of terms such as "consultant," "teacher," and "collaborator" that convey a more active role for the community in the research. Czaykowska-Higgins (2009) builds on Cameron et al.'s (1992) empowerment research to propose an additional model for engagement, Community-Based Language Research, and adding to the roles of community members in the research "co-investigator" and

"partner." Grinevald (2003) also adds *by* to Cameron et al.'s (1992) empowerment model and suggests linguists should consider that "sometimes doing no fieldwork on an endangered language is best" (pg. 62). Citing cases of community-internal division on the perceived benefits of the research, Grinevald advises linguists to go against the "hype campaigns" for "'saving' endangered languages" and "some absolute value of science" to reflect upon the potential impacts "before embarking into some field situation" (pg. 60-61). Despite recognition of the normative ethics and the scientific colonialism motivating some work in linguistics, this argument still frames the linguist as the decision maker in this process.

Leonard (2017) argues for actively facing the colonial past that linguistics is built on and a process of decolonizing linguistics by making room for community definitions of language and language work, including research, teaching, and advocacy. He advocates for a rights-based model of language work that focuses on the community's reclamation of their language and cultural practices rather than the current model of language revitalization that focuses on the language itself and increasing the number of speakers through a decontextualized learning pedagogy. Even in collaborative projects built on respectful linguist-community relationships, entrenched colonial methods and understandings can result in research products that are misaligned or even unusable for community needs (Leonard, 2021). Leonard (2021) suggests "the sharing of knowledge rather than information, an emphasis on positionality and reflexivity, and the consideration of all relationships when identifying the stakeholders in a given project" as ways to uphold indigenous epistemologies and create linguistic research that moves away from its colonial foundations (pg. 28). Here, Leonard draws on Smith's (2012) distinction between knowledge and information, where information refers to only basic facts or conclusions and knowledge includes analyses and theories for deeper understandings and representations. Tsikewa (2021) builds on the call for linguistics to address its colonial foundations and proposes improvements to the linguistic field method curriculum so that future linguists may be more aware of and attentive to indigenous epistemologies and methods.

More linguists have begun to directly bring attention to the racism and colonialism rooted in the field of linguistics and related technical fields (Bird, 2020; D'Arcy and Bender, 2023). Pointing out the dearth of linguists of color entering the field, Charity Hudley et al. (2020) call for a reevaluation of how linguistics incorporates interdisciplinary approaches to race: "we must draw upon and contribute to scholarly theories of race by asking what methods will best enable us to be racially inclusive in our work as well as by exam-

ining language and race in ways that aim for social justice" (Charity Hudley et al., 2020, pg. e219). Leonard (2020) and Gaby and Woods (2020) each respond to this call, diving into the specific implications of such a reevaluation for indigenous communities and advocate for the prioritization of indigenous communities' linguistic rights and goals in linguistic research. Gaby and Woods (2020) recommend several steps for linguists to support social and racial justice for indigenous peoples, including rep/matriating[1] indigenous language knowledge, recognizing indigenous knowledge sovereignty, avoiding degrading and dehumanizing language and rhetoric in reference to indigenous peoples, and acknowledging the social and political context(s) of language. Focusing on the indigenous experience in the U.S., Leonard (2020) points out that Native Americans are in fact a political group, rather than a race; the racialization of Native Americans through, for example, the use of blood quantum as a citizenship prerequisite deliberately undermines the status of tribes and groups as sovereign nations (see also TallBear, 2013). In addition to supporting Charity Hudley et al.'s (2020) call to engage with interdisciplinary scholarship that centers the marginalized, Leonard (2020) highlights the intertwined factors of racism and colonialism by emphasizing the need for linguistics organizations to recognize both that Native American is a political category and that Native American nations can and do reinforce racist and exclusionary policies.

2.1.4 Indigenous-Led Research

An underlying theme throughout indigenous research is self-determination: indigenous researchers using indigenous methods to ask questions that support indigenous communities' goals. As Cram (2001) states in arguing for indigenous research localized for the people of Aoteoroa (also known as New Zealand), "Māori research by, with and for Māori is about regaining control over Māori knowledge and Māori resources" (pg. 37). Cram (2001) discusses seven principles across prior work in Kaupapa Māori (that is, Māori principles and philosophy) as a framework and methodology for achieving research by, with and for Māori. This is not to say that collaboration between indigenous and non-indigenous partners is discouraged; collaboration in which indigenous partners are truly equal and contribute to decisions made at all states of the work including the design and implementation of project's aim, methodology, and the outcomes are welcomed

[1]Repatriation may refer to the return of human remains, artifacts, and/or cultural knowledge to descendant indigenous communities; rematriation is also used to counter the heteropatriarchal assumptions that are embedded within the term repatriation, especially by communities with matrilineal social practices (see Tuck, 2011, for further discussion).

(Leonard and Haynes, 2010; Yua et al., 2022). What the appropriate methods are for each community is dependent on the unique context of that community, with a growing body of work presenting localized research methodologies (Smith, 2012; McCarty et al., 2018; Chilisa, 2020; George et al., 2020). Leonard (2017) illustrates how indigenous understandings of key terms in linguistic research, such as "language," and methods may be articulated by various community members in three case studies on language reclamation. These fundamental differences show how vital it is to attend to the unique perspectives of each indigenous community. The strength in pushing for self-determination allows for disparate indigenous communities to come together for political power while still respecting and maintaining these individual perspectives.

Ethical research guidelines have emerged as a tool for indigenous communities to communicate their perspectives and values with respect to research, giving indigenous communities more control over what research is conducted within their communities and by whom. Building off of Cram (2001), Ormond et al. (2006) propose protocols for working with Māori communities based on three emerging ethical principles: relationships between researchers and communities; researchers knowing themselves and their positionality within the research context; and the aspects of research done the Māori way that help protect the community. Hudson et al. (2010) also reference this work in establishing guidelines for researchers as well as recommendations for community members who represent Māori communities on ethical review boards. In indigenous North America, many tribal research protocols and review boards are founded on the Four Rs of indigenous research: *respect*, *relevance*, *reciprocity*, and *responsibility*. First put forward by Kirkness and Barnhardt (1991), Tsosie et al. (2022) later expanded the four Rs to six, adding on *relationship* and *representation*. The Indigenous Data Governance (CIDG)'s Indigenous Ethics Guiding Documents List[2] presents these ethical research guidelines and nearly 30 others, largely localized to North America as well as others from Oceania, Africa, and South America (David-Chavez et al., 2020; Natonabah et al., 2020).

With the establishment of ethical review boards and community research guidelines across geographic regions, meta-analyses of the established boards and guidelines have also been conducted. Tunón et al. (2016) take a broad approach and analyze 12 ethical guidelines for biodiversity developed by Saami, First Nations, and Māori communities as well as the United Nations and academic associations and institutions from Austalia, Canada, Sweden, and the United States. They compare the values in ethical codes and guide-

[2]Available at `https://indigenousdatalab.org/indigenousdatastewards/`

lines related to indigenous knowledge and cultural practices from those different authoring institutions to understand the development processes of those codes and guidelines and study researchers' awareness of them. Using the 18 principles of the International Society of Ethnobiology (ISE) Code of Ethics (International Society of Ethnobiology, 2006 with 2008 additions) as a point of comparison, they find 5 "core" ethical principles across the guidelines: respect, recognition of rights, responsibility as a scholar, participation, and mutual benefits. They concluded that differences between the guidelines arose from the various contexts in which they were developed and conflicts between the guidelines arose from the differing author perspectives. Kuhn et al. (2020) and Hayward et al. (2021) explored more localized contexts in their analyses. Kuhn et al. (2020) analyzed tribal institutional review boards in the U.S. for common structures and principles. Searching through health and human research protections databases, they selected six review boards for analysis and surfaced similarities across the administration, research application processes, and research management requirements of the selected boards. They also found four common guiding principles: honoring tribal sovereignty, the advancement of indigenous research ethics, research benefits at all levels within the respective community, and a commitment to protecting natural resources for both present and future generations. Hayward et al. (2021) performed a similar analysis, asking three research questions related to ethical research guidelines from First Nations, Métis, and Inuit communities and organizations in Canada: 1) what are the existing guidelines and protocols in Canada, 2) how are indigenous values integrated into these guidelines and protocols, and 3) how have research processes and outcomes been shared by these guidelines and protocols? They surveyed 20 indigenous communities' research guidelines, collecting information where available on the processes that led to establishing those guidelines, and showed the impact those guidelines have had on research approaches in academic and government institutions. Hayward et al. report on three surfaced themes across the guidelines: balancing individual and collective rights; upholding culturally grounded ethical principles; and self-determined research processes, methods, and knowledge translation. They call for Canadian research institutions and local indigenous peoples to "work collaboratively and creatively" towards new research processes based on these principles (Hayward et al., 2021, pg. 414). Throughout these approaches, recognition of indigenous peoples' rights to determine research methods and benefits to the community reoccur. While these surveys and similar work present other indigenous communities with possible example research guidelines and protocols and analyses to learn from in devel-

oping their own frameworks, surveys across varied communities must be careful to not replicate colonial practices by, for example, essentializing categories or creating hierarchies of communities or values.

In an increasingly digital world, Indigenous Data Sovereignty (IDS) has emerged as a new framework for protecting indigenous communities' rights and ownership over their knowledge, data, and digital resources (Carroll et al., 2019; Walter et al., 2021). In this framework, data is argued to be central to modern indigenous governance and policy-making. Furthermore, keeping data ownership localized to the community prevents others from using the data to misrepresent the community or reproduce colonial data practices. Organizations such as Te Hiku Media,[3] a Māori-owned organization dedicated to Māori language reclamation, and the Exchange for Local Observations and Knowledge of the Arctic (ELOKA),[4] a collaborative center between indigenous communities in the Arctic and the University of Colorado Boulder for environmental stewardship, operate under data sovereignty frameworks. Many indigenous communities, however, are still building their capacity and infrastructure for digital data collection, storage, and maintenance. The First Nations principles of OCAP® were among the first standards for indigenous data management that emphasized indigenous Ownership, Control, Access, and Possession of indigenous data in Canada (First Nation's Information Governance Centre, 2018). Carroll et al. (2020) developed the CARE principles[5] to bring indigenous communities' needs into the conversation around international open data policies. Designed to work in concert with the FAIR principles of Findable, Accessible, Interoperable, and Reusable data (Wilkinson et al., 2016), the CARE principles of Collective benefit, Authority to control, Responsibility, and Ethics assert indigenous communities' sovereignty over their own data in an information ecosystem that otherwise holds universally available data as the ideal. Other tools such as the open-source platforms Indigitization[6] and Mukurtu[7] provide indigenous communities with customizable infrastructure for data storage and protected access, while the Local Contexts Traditional Knowledge (TK) Labels[8] ensure communities are acknowledged for their contributions to the data and enable appropriate handling of indigenous data (Christen, 2015). These tools and frameworks support indigenous communities in realizing their own goals for data governance and for their futures.

[3] https://tehiku.nz/
[4] https://eloka-arctic.org/about-eloka
[5] Available at https://www.gida-global.org/care
[6] https://www.indigitization.ca/
[7] https://mukurtu.org/
[8] https://localcontexts.org/labels/traditional-knowledge-labels/

2.1.5 Ethical Considerations for Indigenous Data and Language Technology

Developers of language technology for indigenous communities need to be wary of reproducing colonial practices in digital forms. Grinevald (2003) warned against hype campaigns and the illusion of objective science; these warnings continue to be relevant when considering technology. Technological development for indigenous communities must be directed by and with those communities in alignment with their needs and goals, not the needs and goals of other societies or a supposedly disembodied science. The United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) (UNDRIP) makes explicit indigenous peoples' rights to self-determination, to free, prior and informed consent, and to control the development of technologies based on their traditional knowledges and languages (Office of the High Commissioner for Human Rights, 2007).

In a similar vein, the warnings against deficit characterizations of indigenous communities and their languages still apply to technology (Gaby and Woods, 2020; Leonard, 2020). In NLP, researchers tend to characterize languages by relative number of machine-readable (mostly text-based) resources that are available in that language, with a handful of national languages and especially English being described as "high-resource" and all others termed "low-resource". Text is the primary format for language in many technologies, and even audio processing often involves converting the wave form to standardized text prior to further processing. Bird (2022) points out that the characterization of "low-resource languages" is a colonial framing that fails to take into account the diversity of use of these languages within this category and the other digital, physical, and human localizations of language knowledge. Bird (2022) and Liu et al. (2022) push back against the Western insistence on writing systems and language standardization: "For these communities, linguistic variation is not a problem to be solved but an important element of a vital language ecology" (Liu et al., 2022, pg. 3937). Both point to communities expressing the need for culturally appropriate pedagogical material that supports community language learning and some have leveraged technology to support language learners (e.g., Frey, 2020, for Eastern Cherokee).

Some communities also choose not to collaborate on or engage with technology for a variety of reasons. Contrary to the narrative that "endangered languages must be saved at all costs," "Indigenous people have the option to abandon their languages" and they have the option to refuse technology as the vehicle for any language activities or pedagogy (Bird, 2020, pg. 3507). Technology also opens indigenous commu-

nities to new forms of digital surveillance (Zaugg, 2019). The United Nations Educational, Scientific and Cultural Organization (UNESCO)'s Action Plan for the International Year of Indigenous Languages (IYIL) listed among the goals for the 2019 effort the "preservation of indigenous languages, access to education, information and knowledge in and about indigenous languages for indigenous children, young people and adults, improvement of data collection and sharing of information in and about indigenous languages, using language technology and other communication and information mechanisms (access)" but notably excludes any guarantee that indigenous communities will maintain control or ownership of that collected data and any access protocols (Secretariat of the Permanent Forum on Indigenous Issues, 2018, pg. 8). Communities may also reject technology as a means for rejecting the capitalistic systems that technology is embedded in. Although benefit sharing agreements can and have resulted in significant economic prosperity for several indigenous communities (National Khoisan Council & Cederberg Belt Indigenous Farmers Representatives, 2019; Argumedo et al., 2011), these agreements "remain dependent on a predatory economy that is entirely at odds with the deep reciprocity that forms the cultural core of many Indigenous peoples' relationships with land" (Coulthard, 2014, pg. 171). Less than equitable mutual economic benefits are particularly offensive as they re-inflict colonial exploitation on indigenous peoples. When Lion Bridge, an American language technology localization company, approached Māori language groups asking them to provide audio data of Māori for $45 (US) an hour, Te Hiku Media published their rejection of the offer to other Māori groups and urged them to reject the offer as well (Coffey, 2021). Peter-Lucas Jones, the CEO of Te Hiku Media, called out the injustice of providing labor and language data to a company that would then profit on making Māori available through third party technology: "They suppressed our languages and physically beat it out of our grandparents... And now they want to sell our language back to us as a service." Accepting the offer "would mean that Māori would miss out on the economic opportunities created using the language that belongs to them, much like they didn't see the economic benefits of the land that belonged to them" (Coffey, 2021, n.p.). Supporting indigenous self-determination and autonomy also means supporting indigenous refusal.

2.1.6 Summary of Indigenous Data Governance and Ethics in the Context of Language Technology

In this section, I summarized the role that research played in subjecting indigenous peoples to colonial violence (§2.1.1). This role has changed over time as indigenous communities have asserted their rights in government policy (§2.1.2). Because of the efforts of indigenous communities and allied researchers, communities may now work to challenge the colonial foundations embedded in language research (§2.1.3) and lead their own research agendas (§2.1.4). Future technical research with indigenous communities can avoid repeating past colonial harms by prioritizing communities' rights to self-determination, avoiding deficit characterizations of indigenous communities and their languages, and respecting their right to refuse collaboration or technical solutions (§2.1.5). Research driven by, developed for, and conducted with indigenous peoples presents the opportunity for indigenous communities and their collaborators to re-imagine and co-develop creative tools rooted in indigenous histories "to promote intergenerational transmission of knowledge, ceremony, and practice, to connect and enhance our communities and to frame our relationships to the land, sea, and skyscapes" (Lewis et al., 2020, pg. 20). In the next section, I explore the ways in which the progress towards indigenous self-determination has influenced signed language communities in advocating for their linguistic rights and the similarities and differences they face with respect to language technology development.

2.2 Signed Languages in Linguistics and NLP, Deaf Culture, and Policy Making

Linguistic research on signed languages began in the 1950s (Tervoort, 1953) and 1960s (Stokoe, 1960). These works established signed languages as natural languages, each distinct from any spoken language and in equal standing with spoken languages in terms of their complexity and structure. Research on signed languages now includes the development of national corpora like the Australian Sign Language corpus (Johnston, 2009), the British Sign Language corpus (Fenlon et al., 2015), and the German Sign Language corpus (Hanke et al., 2020); investigations into the varieties and sociolinguistic features within a given signed language (e.g., Hill, 2017); and documenting rural community signed languages, often termed village

signed languages, around the world (Zeshan and de Vos, 2012). Research on deaf communities and signed languages since then has grown to encompass its own interdisciplinary literature as the field of deaf studies (see e.g., Kusters et al., 2017b).

Deaf communities are still working to codify their rights to their signed languages and to access through signed languages. I provide a brief overview of the history of discrimination against deaf peoples and the implications this history has for education and language policy, particularly in the United States (§2.2.1). This overview is by no means comprehensive but rather is intended to provide a starting point for considering the connections to other histories of oppression (§2.2.2), the present struggles for signed language rights (§2.2.3), and the possible paths forward in technological innovations for signed languages (§2.2.4). signed language processing (SLP) technologies for ASL and other signed languages have been limited in the past, however recent advances in computer vision have led to renewed calls for SLP technologies (Yin et al., 2021; Bragg et al., 2019). In order to appropriately evaluate these systems, developers need to understand deaf history and culture and the implications for sign language and deaf policies within their local contexts. Technologies designed for, with, and by deaf peoples will best address the needs of deaf communities; §2.2.5 highlights some of the ethical considerations specific to working with deaf communities on SLP and signed language dataset development.

2.2.1 Hegemony of Spoken Languages

Generally, children acquire one or more languages from their parents and guardians in their first few years of life in a process linguists call intergenerational transmission. Children, both deaf and hearing, whose families frequently use one or more signed languages will acquire these languages over time as the children learn from and engage with their parents or guardians. Deaf children, however, are more often than not born into hearing families who are not already proficient in a signed language (Johnston, 2006). Without some kind of early intervention, deaf children born to families who use spoken languages do not have access to a language they can easily acquire. Significant adverse health outcomes have been documented in children who have suffered language deprivation, meaning they are not provided sufficient access to a natural language within the first few years of life (Murray et al., 2019). Language deprivation causes cognitive development delays and difficulties with socioemotional development due to the inability to communicate and build relation-

ships with family and peers. Systematic language deprivation has contributed to the false stereotype of deaf people being "less than" throughout history (Lane, 1984). Deaf people have been subjected to paternalistic ideologies that frame them as incapable when in fact societies built for spoken languages isolate deaf people from the surrounding community and restrict their access to information and education (Padden and Humphries, 1988).

Many deaf people learn signed languages in public deaf spaces such as schools rather than in the home (Ladd, 2003). Traditionally, deaf schools were established by clergy such as Lauren Clerc and Thomas Gallaudet who established the American School for the Deaf in the United States in 1817 using signed language and pedagogical methods from France. Signed language pedagogical methods continue to be used in bilingual-bimodal deaf education, where students first learn the local signed language and then learn the written form of a spoken language as a second language. This form of education fits within the *sociocultural model* of deafness, where deaf identity is a cultural and linguistic identity. Humphries et al. (2014) advocate for all deaf children learning a signed language, regardless of choices made regarding assistive hearing technology and oral training, to ensure early acquisition of a language and because being bilingual has cognitive, social, and educational benefits.

However, many factors led to pushback against using signed languages in deaf education, particularly in the United States and in Europe, starting in the 19th century and continuing through the 20th century. In the United States, spoken English and the majority hearing culture were seen as "normal"; signed language and the development of a separate deaf culture were therefore seen as foreign and undesirable (Edwards, 2012). These sentiments led to the rise of oralism, an education policy in which deaf students are taught spoken language through lip-reading and practicing speech and deaf people are discouraged from interacting with one another (Ladd, 2003; Edwards, 2012). Despite signed languages being used in schools for deaf people for the first half of the 19th century, the 2nd International Conference on Education of the Deaf in 1880, in Milan, Italy, resolved to promote only oralism in education and banned the use of sign languages in classrooms (World Federation of the Deaf, 2019). The prevalence of oralist pedagogy has decreased in the later half of the 20th century as the methodology has been reported to have harmful impacts on deaf students and is generally seen as an unsuccessful educational policy for deaf students (Lane, 1992). Artificial systems for manually coding languages, including several for English such as American Signed English and

Seeing Essential English, have been developed to convey the word order and content of spoken languages in the visual modality. While these systems appear to be a compromise, they still maintain the perceived superiority of spoken languages like English and have been used to discourage the use of natural signed languages like American Sign Language (ASL). Despite decades of research showing that signed languages are fully functional languages in their own right (Tervoort, 1953; Stokoe, 1960), perceptions that signed languages are not "real" languages still exist (Krausneker, 2015).

Oralism and assistive hearing technologies fall under the *medical model* of deafness. In this view, deafness is seen in a deficit perspective as the loss of hearing or inability to hear that may be fixed or mitigated. Lane (1992) accuses research, medical and educational professionals of being economically incentivized to promote cochlear implants and other hearing-aid technologies despite the harms medical interventions inflict on deaf individuals and deaf communities. The advent of medical interventions such as cochlear implants and hearing aids has renewed the emphasis on deaf children learning spoken language and being integrated into hearing society through practices such as mainstreaming (Ladd, 2007). Mainstreaming is an educational policy that places deaf students in mainstream public schools in the name of inclusion, however it can lead to deaf students being more isolated from other deaf peers and research into the educational outcomes of deaf students in mainstream versus segregated deaf educational environments is inconclusive due to compounding social factors (for a comprehensive discussion on this topic, see Mathews, 2018). While deaf advocates largely favor bilingual-bimodal education, there still exists medical research which promotes the use of cochlear implants for deaf children and actively discourages the use of signed language (e.g. Geers et al., 2017). Despite the fact that signed languages are natural languages and have existed alongside spoken languages, and despite the known harms of language deprivation, deaf people are still fighting for their right to education and to signed languages as a medium of instruction.

2.2.2 Parallels between Deaf Studies and Indigenous Studies

Deaf scholars have drawn parallels between the forces of oppression applied to deaf communities and other communities who have been subjected to racism, colonialism, and discrimination (Ladd, 2003). Humphries (1975) adds deaf experiences to this list with the term *audism*, used to describe systematic and individual instances of discrimination and prejudice against deaf peoples and signed languages. The frameworks and

models from disciplines such as Black studies, feminist studies, queer studies, disability studies, and indigenous studies have influenced and been employed in work for researching deaf experiences (O'Brien, 2017). Connections to indigenous experiences have been drawn on the basis of similar struggles for self-determination, linguistic rights, and culturally-appropriate educational policies (Ladd, 2003; Batterbury et al., 2007; Bone et al., 2022). Rather than adapting indigenous frameworks and applying them in deaf studies analyses, O'Brien (2017) instead looks to Kaupapa Māori research to inspire the creation of a deaf-led deaf studies research framework. Following a workshop organized to discuss the topic, O'Brien (2017) presents four possible principles for deaf-led research: the primacy of signed languages, self-determination, identity preservation, and community development. Examining the ways in which these principles align with indigenous issues and where they diverge will help to situate the work in the later chapters of this dissertation.

While the value of self-determination has been a strongly unifying foundation for both indigenous and deaf peoples to build power internationally, the perception of being a singular group can instead result in universal solutions and simplified definitions that in fact reduce individual communities' abilities to determine their own futures. For this reason, Ladd's (2003) use of colonialism as a metaphor for framing the oppression of deaf people has drawn criticism for relying on the use of nationhood and a uniform experience of deafness to establish deaf power (Anglin-Jaffe, 2015). Anglin-Jaffe (2015) argues that deaf peoples' experiences with auditory perception and signed languages vary greatly; trying to define whose experience of deafness "counts" risks alienating people who identify as deaf but do not neatly fall within the defined criteria. In considering how to apply decolonial theories to "support and acknowledge the cultural elements related to sign languages and regional Deaf communities in a way that celebrates and respects them in deaf education through the curriculum and pedagogy, without imposing essentialist, exclusive or elitist practices," Anglin-Jaffe (2015, pg. 91) suggests the use of interdisciplinary studies and genealogy methods. Similarly, in drafting the UNDRIP, indigenous peoples argued against an international definition of "indigenous," insisting that each indigenous people should have the right to define themselves (Asia Pacific Forum of National Human Rights Institutions and Office of the High Commissioner for Human Rights, 2013).

Writing is one issue which has posed challenges for both indigenous communities' and signed languages communities' linguistic self-determination. It is through the written form that most deaf peoples become

bilingual in spoken languages, however there are no widely accepted writing systems for signed languages. Some indigenous languages have their own writing systems, such as the Cherokee syllabary, but many do not and do not see standardized writing systems as a necessary goal (Liu et al., 2022; Bird, 2022). In linguistic analyses of signed languages and indigenous languages, transcriptions of the language data use glossing, or the use of a word in a language of broader communication, to represent a unique sign in a signed language (Hochgesang, 2022) or a morpheme in an indigenous language (see e.g. Lewis and Xia, 2010). For indigenous languages, this means that the basic written representation of the language data is most often the transcription of the sounds in the phrase being analyzed using the International Phonetic Alphabet (IPA) or a linguist-developed practical orthography. However, the basic written representation of signed language data is the glosses from a spoken language, meaning there is no intermediary representation of the linguistic unit(s) within a sign that allow for a written form of the signed language independent of spoken language in the way that IPA allows for written representations of indigenous languages independent of English or any other national language. In both cases the written forms serve as a machine-readable format of the language that provides the input to language technology, enabling search functions for a unique sign or morpheme in a dataset and analysis of the contexts that the sign or morpheme may appear in. Academic publications also tend to be written in English. While dissertations and abstracts may be written in indigenous languages and videos of signed language interpretations of papers may be published along with articles, scholars are still pressured to published in English to meet academic norms and expectations.

O'Brien (2017) describes identity preservation as the feeling that deaf scholars experience in resisting the pressures to conform to academic expectations and maintain connections, time, and resources to pursue deaf values and deaf-focused research. Haualand (2017) similarly recounts her experiences of being an international deaf scholar at Gallaudet University in the United States and compares the perception of being different from others on the basis of nationality to the perception of being different on the basis of deafness at her home institution in Norway. Despite communicating in ASL instead of Norwegian Sign Language, Haualand expresses a greater sense of ease at Gallaudet University where communication in a signed language is the norm and her legitimacy as a deaf academic was not called into question. Leonard (2017) and Hermes (2012) discuss similar pressures in being indigenous scholars and prioritizing research towards community development. That being said, the academy has seen a greater shift towards accepting

indigenous community development as research than deaf community development. For example, in the United States, signed language scholars struggle to get funding for language documentation of signed languages because languages like ASL do not meet the criteria for "endangered" the way indigenous languages do (Hochgesang et al., 2023).

2.2.3 Legal Recognition of Signed Languages

The rights of deaf and deafblind individuals to access and learn signed languages in visual and tactile modalities are largely granted under the medical model of deafness, with examples such as the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) (Office of the High Commissioner for Human Rights, 2006) and the Americans with Disabilities Act (1990). In the United States, ASL is legally recognized by many states, but for the purposes of foreign language credits for university students rather than the education of young deaf children (Murray, 2019). From 2016 to 2020, the number of enrollments in university-level ASL courses increased from about 66,000 to about 69,000 as the third most commonly taught language in US higher education, even as the enrollments for most other world language courses decreased 3%-32% in the same time frame (Modern Language Association, 2022). Deaf people in the United States are still fighting for the constitutional right to sign language for deaf children and their families (Humphries et al., 2013).

Many deaf communities around the world are building political movements to fight for legal recognition of their linguistic rights. However, because of the competing models of deafness, deaf signers sometimes have legal status as people with a disability, sometimes as peoples who use minoritized languages, and often both. Murray (2015) uses this variation in legal perception of deaf signers to categorize the existing laws recognizing signed languages into explicit recognition and implicit recognition (see also De Meulder et al., 2019b). The Americans with Disabilities Act is an example of implicit recognition, where ASL is implicitly recognized in terms of disability access. Batterbury (2012) argues the UNCRPD may still be a useful policy tool for signed language rights despite its basis in the medical model of deafness; while it frames deafness in terms of disability, it acknowledges signed languages explicitly and includes a mechanism for monitoring states' progress towards its stated goals. The adoption of the UNCRPD in 2006 has led to more countries using disability rights as the basis for signed language recognition, including in Korea, Chile, and Malta and

in ongoing debates in the Netherlands and Italy (De Meulder et al., 2019b).

Other states have explicit legislation recognizing signed languages in dedicated language laws. Some examples include the Catalonian LSC (Catalan Sign Language) Law, the New Zealand Sign Language Act, and the Finland Sign Language Act. In these cases, the legislation was influenced by or based on prior minority or indigenous language legislation. The Catalonian LSC (Catalan Sign Language) Law was based on the Aranese language minority law; the New Zealand Sign Language Act was influenced by the Māori Language Act (which itself was influenced by the Welsh and Irish language acts); and the Finland Sign Language Act was based on the Sami Language Act (De Meulder et al., 2019b). Cokart et al. (2019) discuss 30 years of mobilization and political activism for the legal recognition of Dutch Sign Language (NGT). As of October 13, 2020, the Netherlands has officially recognized NGT as a national language alongside Dutch and Frisian.

Even with explicit recognition, however, signed languages are often not afforded the same resources as spoken languages, particularly in education. International signed language communities, led by the World Federation of the Deaf (WFD), advocate for moving away from deaf rights as disability rights, which accord rights to individuals, and instead towards rights accorded to the group as a cultural and linguistic minority. The WFD Charter on Sign Language Rights for All states: "We emphasise the paradigm shift from the medical model of disability to the human rights model of disability in line with the UNCRPD. Deaf people are human rights holders entitled to equal opportunities to participate in society in the same way as other citizens" (World Federation of the Deaf, 2019, p.g. 1). The Charter stresses access to these equal opportunities "through the use of sign languages" and, similar to the indigenous sovereignty movement, also argues for deaf self-determination such that deaf people are "actors of their own destiny and inclusion in society" (World Federation of the Deaf, 2019, p.g. 1). Further work is needed to ensure that legal recognition of signed languages leads to early access to signed language education for deaf children and their families and broader inclusion of signed languages in public spaces.

2.2.4 Current State of Signed Language Technology

New processing capabilities, particularly in the field of computer vision, have greatly increased interest in signed language technology from technical fields and academics. Yin et al. (2021) provide a selection of

technical processing frameworks (usually called *tasks*) that are commonly investigated in SLP. Detection systems take as their input a video frame and output whether or not a signed language is being used within that frame, in terms of either true (signed language is being used) or false (signed language is not being used). Signed language identification systems classify a given video by the specific signed language used in the video, such as ASL or Brazilian Signed Language (Libras). Segmentation technologies are designed to label the boundaries of meaningful units of signs or phrases in a given video. Signed language recognition systems can vary in terms of whether they take an image or a video as input. In the case where the input is an image or a video with a single sign, the system will output a label indicating the gloss for that sign. In the case of a video with continuous signing, the system will output segmented video with a gloss as a label for each video segment. What is frequently called *signed language translation* tends to refer to only systems that take as input videos of signed language and output a spoken language interpretation. Systems that take as input text from a spoken language and output signed language video or animations are termed signed language production. Across all of these tasks, Yin et al. (2021) argue for more linguistically informed design of SLP systems.

Bragg et al. (2019) also argue for more interdiscplinary research for SLP as well as larger and more representative datasets designed to model real world domains. Currently many signed language datasets are designed for either linguistic research or machine learning research with little overlap in the structure or annotation of the datasets (De Sisto et al., 2022). Signed language datasets are also much smaller than text-based or even speech-based datasets for spoken languages (Forster et al., 2010); in such cases, as with smaller indigenous language datasets, uncritical applications of machine learning techniques designed for larger amounts of data tend to produce incorrect or even harmful results. Signed language data, particularly ASL data, is becoming more broadly available (see e.g., Hochgesang et al., 2017-2022; Dudis et al., 2020; Desai et al., 2023). However, with this broad availability, it is increasingly important to collaborate with signed language communities and address problems they identify as suited to technical solutions.

There are longstanding concerns about technologists conducting research on signed languages without learning about signed language communities prior to the work (Braffort, 2002). Signed language gloves, for example, are one application of SLP that have not been designed to serve signed language communities (Erard, 2017; Hill, 2020). Developers of these tools (often hearing individuals) claim the technology, worn on

the hands of signers, are tools for translation. However, this claim is false, as the gloves often only work for finger-spelling, and even if they could process lexicalized signs, they would miss all the non-manual features that are required for interpreting signed language (Forshay et al., 2016). Still further, the intended purpose of signed language gloves is to facilitate hearing individuals understanding signed language without providing the signer access to spoken language, placing all of the burden for communication on the signer without any reciprocated effort from the hearing individual. Signed language avatars have also caused concern, especially when the technology is intended to be used in high-stakes interpretation situations like medical contexts. This work is done without critical understanding of signed language interpretation rights and access, and is often trained on data from signed language interpreters (De Meulder, 2021). Framing signed language interpreters as the ideal signer both misunderstands the problem (signed language interpreters are not the intended audience for avatars and are not broadly representative of any signed language) and, as De Meulder (2021) points out, leaves unquestioned the assumption that the political institution of signed language interpretation services is a model that should be replicated, with technology or otherwise.

2.2.5 Ethical Considerations of SLP

With the pitfalls of SLP in mind, Prietch et al. (2022) call for "more inclusive and qualitative research for, with and by Deaf persons who are sign language users" (p.g. 1). They analyze 37 research studies in SLP and find that only two used a design developed by deaf participants and only one paper discussed the cultural aspects of the intended deaf community in the design of the system. This echoes Bragg et al.'s (2021) call for researchers to build trust with deaf communities and include deaf individuals in studies and as members of design teams. They suggest that hearing researchers work to collaborate with existing deaf research groups and advocate organizations in order to reduce concerns about cultural appropriation of signed language data. However, their framing of trying to attract the interest of technical communities through competitions and ML-friendly datasets, which take significant time, effort, and resources to create, leaves unchallenged the power that technical communities have with respect to signed language communities. These considerations of power are especially critical when collaborating with signed language communities in more vulnerable contexts (Clerck and Lutalo-Kiingi, 2018; Cooper and Nguyễn, 2015; Braithwaite, 2020; Hochgesang, 2015). Framing collaborations as interdisciplinary exchanges where each collaborator has something to

teach the other and contribute to the research lends itself to a more equitable collaborative space.

Bragg et al. (2021) provide a comprehensive overview of signed language datasets and their ethical considerations when used in an ML context. These considerations include data collection, ownership, and access. The method for collecting data, whether data is elicited or found, will always have an impact on the resulting dataset. In cases of data elicitation, the data collector's identity can impact the collection effort as signers shift their signing style to accommodate their interlocutor and the collection methodology. Although signed language data may be available online, platforms may or may not have terms related to collecting the data, and efforts should be made to getting the video owners' consent (see also Hou et al., 2020). In discussing ownership, Bragg et al. (2021) illustrate the many ways that data ownership may be realized: physical ownership, legal ownership, monetary ownership, cultural and linguistic ownership, and perceived ownership. Each of these have implications for the responsibilities that the owner of the data has. One of these responsibilities is protecting the privacy of the people represented in the data, either through consent processes or through anonymization approaches (Isard, 2020; Bragg et al., 2020). Providing levels of access to the data, and ensuring that the responsibility for abiding by those access levels is transferred to anyone granted access, is another data ownership responsibility. Schulder and Hanke (2022) show how the German Sign Language (DGS) corpus was developed in line with the FAIR principles (Wilkinson et al., 2016) and the CARE principles (Carroll et al., 2020) with considerations for the collection process, informed consent in German and DGS, public and protected access, privacy for participants and individuals discussed in the data, and participant attribution available in the protected metadata.

Emphasized throughout deaf-led SLP work is that work with signed language communities often requires work to be done *in* signed languages as the community's preferred method of communication. This includes consent processes, communication with participants and team members, and in dissemination. Hochgesang and Palfreyman (2022) discuss the ways in which signed languages have been analyzed in research through the lens of spoken language words (often English) and advocate for signed languages being presented as much as possible in the form of videos and images in addition to glosses and translations.[9] Even committed efforts towards signed language communication have to rely on text at times however. While the

[9]They attribute the Twitter hashtag #GlossGesang to Carl Börstell: "Always present sign language data in a visual format (videos/images) without relying solely on glossing". Available at `https://twitter.com/c_borstell/status/1177498599992610823`.

web portal for the ASL Citizen dataset (Desai et al., 2023) provides information in the form of both text in English and videos in ASL for the overview and description pages, it does not provide videos in ASL for the pages with the license and completed datasheet (Gebru et al., 2021). As direct stakeholders of the dataset, it is important for the information in the license and datasheet to be accessible to deaf people; however, how best to communicate long-form technical and legal information in signed language videos remains an open design question for signed language communities and their collaborators.

2.2.6 Summary of Signed Languages in Linguistics and NLP, Deaf Culture, and Policy Making

I began this section with the history of discrimination against deaf peoples, especially in the United States, with a focus on the denigration of deaf people for their use of signed languages (§2.2.1) Drawing on analogous experiences of linguistic oppression, deaf studies scholars have at times applied theories and analyses from other fields of study centered around peoples who have suffered oppression, including indigenous peoples around the world, in developing and growing their own body of research (§2.2.2). Ongoing efforts towards legal recognition of signed languages have also found inspiration in the legal recognition of indigenous languages in the nations where they exist, though other legal frameworks for recognizing signed languages rely on legislation for disability rights (§2.2.3). In addition to increased attention for signed language legislation, recent advances in video processing technology have also led to more interest in the technical aspects of SLP (§2.2.4). However, these technological approaches to signed language processing tend to suffer from a lack of understanding of signed languages, signed language communities, and their cultures and therefore fail to address the needs of deaf peoples. More deaf-led collaborations are needed to address the ethical concerns throughout the stages of technical development, including design, data collection, system building, evaluation, and public release (§2.2.5). The ethical considerations for SLP technology in many instances highlight the signed language-specific considerations of general ethical concerns for language technology and NLP, discussed further in the next section.

2.3 Ethical Considerations in Natural Language Processing and Machine Learning

Longstanding concerns about bias in technology were given new urgency with the rise of Big Data and the proliferation of ML models trained on that data. Even before ML algorithms became widely embedded in our society, Friedman and Nissenbaum (1996) presented a framework for understanding the sources of bias as realized by technical systems systematically and unfairly discriminating against individual and groups of people. These sources are preexisting (from social institutions, practices, attitudes), technical (from technology constraints or considerations), and emergent (appearing in contexts of use). Friedman and Nissenbaum (1996) suggest mitigation strategies for these potential sources of bias, including considering preexisting bias from the beginning of the development process; envisioning the design, algorithms, and interfaces in use to surface technical bias; and communicating the limitations of a system's potential appropriate use cases. Recent work in bias in ML and NLP has largely focused on technical bias (see Mehrabi et al., 2021, for an overview) and struggled with articulations of preexisting and emergent bias (Blodgett et al., 2020).

Where discussions of preexisting and emergent bias do happen, the blame is usually placed on the training data for the system. Slota et al. (2020) conducted 26 semi-structured interviews related to the social consequences of AI development with stakeholders from AI research, governmental and organizational policy, and legal research and practice. They report on one of their primary findings, that many of their interviewees attributed the failures and successes of AI to data quality issues. Slota et al. found that the "outcomes of AI were seen as escaping the control of the designers due to incomplete, biased, or inaccurate data," (pg. 6). However, Slota et al. disagree with their interviewees, noting that even if the data were somehow complete, unbiased, and accurate, there may still be ethical concerns with the resulting system. They instead argue for better infrastructure for evaluating the impacts of AI and conclude that, "Understanding how some outcome came to be requires an understanding of the full lifecycle of the technology that gave rise to the outcome, from data collection, curation and selection, all the way to how that system comes to be represented and understood in the media" (pg. 9). This conclusion, however, does not address why developers might attribute the negative outcomes of *AI* systems to the training data.

Raji et al. (2021) locate the origins of developers' perceived lack of agency when contending with data

in the exclusionary practices of ethical pedagogies in computer science (CS) departments. These pedagogies focus on transferring technical skills centered around mathematics, logic, and programming expertise and eschew skills from social science and humanities pedagogies such as the critical analysis of data within a particular social and historical context. CS curricula then promote technology as the ultimate fix to entrenched social problems, a belief known as *technosolutionism*, without actually teaching students how a problem may be scoped and framed such that a technical implementation is ethically appropriate, can be used in the intended social context, and sufficiently addresses the motivating problem. Raji et al. argue that interdisciplinary collaboration would not only support identifying social problems suited to technical implementations, but in fact these problems require diverse perspectives and methods to fully comprehend the complexity of the context and possible paths of inquiry.

> "The field of CS has not yet come to the full realization that it deals with problems which exceed its traditional field of competency and in fact its problems are not merely technical or moral problems, but they are in fact transversal problems which require a diverse set of skills, and methodologies. A transversal problem is distinct from an interdisciplinary problem as its solution is not found in-between given disciplines but should be constructed from the effects on the stakeholders that could be or were impacted by it, and from a critique of the types of formal and substantive assumptions, choices, requirements and methodologies that are currently built into AI ethics pedagogy." (Raji et al., 2021, pg. 523)

To create systemic change in the field of CS and counter technosolutionism, Raji et al. (2021) recommend teaching how to collaborate with interdisciplinary colleagues, how to frame problems using a variety of methodological tools and knowing which analytical approaches to use, and how to learn from stakeholders.

Looking to inherently interdisciplinary fields such as science, technology, and society studies, we find that critical work provides, not solutions, but questions that help frame how to engage with, analyze, and find agency in interrogating Big Data. Birhane (2021), in arguing for relational ethics in ML and data science, posits that the belief that bias can be corrected for technically "relegates deeply rooted societal and historical injustices, nuanced power asymmetries, and structural inequalities to mere datasets" (pg. 6). Instead, to center disproportionately impacted communities as the authorities in negotiating just paths forward, she poses questions such as "how might a data worker engage vulnerable communities in ways that surface

harms, when it is often the case that algorithmic harms may be secondary effects, invisible to designers and communities alike, and what questions might be asked to help anticipate these harms?" and "how do we make frictions, often the site of power struggles, visible?" (pg. 6). To situate the frictions within NLP, I summarize the history of ethical considerations in NLP (§2.3.1). These considerations touch on many of the topics discussed in §2.2.5, though the perspective is broadened to general ethical considerations in NLP rather than those specific to signed language communities. I then describe how accountable data practices stemming from ethical considerations have found purchase through documentation (§2.3.2).

2.3.1 Growing Awareness of the Societal Impacts of NLP

Considerations of the ethics and societal impacts of NLP applications began from discussions of the impacts that the development and use of language technology may have on people who are not themselves directly using the technology. Fort et al. (2011) raised concerns about the increasing use of Amazon Mechanical Turk for data creation and annotation and the potential exploitation of crowdsourcing workers. Questions around the ethical practices of the field in using this service led Couillault et al. (2014) to assess the state of documentation of highly used data resources in NLP and found that information on licenses, worker conditions and contracts, and quality assurance details was available for fewer than half of the resources. Around this same time, Sweeney (2013) investigated the language used in targeted online ads that suggested the existence of arrest records for names in Google search queries. She found that not only were ads for public records on a person more likely to be shown for those with Black-associated names than white-associated names, those ads were also more likely to use the word "arrest" in searches of Black-sounding names than white-sounding names. Sweeney (2013) motivated this work through the potential for employment discrimination; for example, if an employer searches for the name of a job applicant and sees the ad suggesting the applicant has an arrest record, they may choose not to hire the applicant, even if no such record actually exists. Her work also revealed implications for the tangible societal impacts of ML and NLP systems.

These findings sparked a number of introspective discussions about NLP's relationship with ethics beyond academic integrity. Following the organization of a Traitement Automatique des Langues Naturelles (TALN) workshop on ethics and NLP, Fort and Couillault (2016) conducted two surveys, one in July 2015 in the French NLP community and one in September 2015 in the wider international NLP community. The

responses from both surveys indicated that researchers were both aware of the ethical implications of their work and interested in exploring broader impacts, with 59.8% of French respondents and 77% of international respondents agreeing that ethics should be included in the list of subjects in the calls for papers of NLP conferences. The first Association for Computational Linguistics (ACL) Workshop on Ethics in Natural Language Processing (EthNLP) was held at the European Chapter of the ACL in 2017 (Hovy et al., 2017). The goal of the EthNLP workshop was to define and raise awareness of ethical considerations in NLP, but it did so by inviting collaboration with and drawing from other disciplines. Several conferences organized around NLP and ML topics, including ACL, the Conference on Neural Information Processing Systems (NeurIPS), and the International Conference on Machine Learning (ICML), have since instituted ethical considerations as a recommended section in submitted papers and ethics reviews as part of their peer review processes.

To support discussions of specific social impacts, Hovy and Spruit (2016) presented a typology of harms caused by NLP technology along with examples for each category. This typology of harms included exclusion, overgeneralization, bias confirmation, topic under- and overexposure, and dual use. Building on typologies of harms from general ML (Barocas et al., 2017; Crawford, 2017), Blodgett (2021) and D'Arcy and Bender (2023) further catalogued the growing number of documented harms caused by NLP technologies. The typology given in Barocas et al. (2017) differentiates between allocational harms, or harms due to the unfair distribution of resources, and representational harms, or harms that contribute to systems of oppression. Blodgett (2021) gives additional examples of these types and introduces a typology of representational harms, consisting of alienation, quality of service, stereotyping, denigration and stigmatization, erasure, and public participation. D'Arcy and Bender (2023) present the harms of NLP technology through the lens of value sensitive design (Friedman et al., 2006b; Friedman and Hendry, 2019), categorizing them in terms of direct and indirect stakeholders (see §2.4 for further discussion of value sensitive design and stakeholder analysis). Within the category of direct stakeholders, there are harms to tech users who use the tech by choice, to those who use the tech not by choice, and to tech developers such as annotators or crowdworkers. Within the category of indirect stakeholders, there are harms to individuals, harms to communities, and harms to those who have unknowingly contributed data to a system or dataset. Both of the surveys presented in Blodgett (2021) and D'Arcy and Bender (2023) show the breadth of harms on marginalized communities

by a wide variety of NLP applications, such as automatic speech recognition systems having higher phonetic error rates for non-white than white speakers of English (Wassink et al., 2022), search engine queries reflecting negative stereotypes of Black women (e.g., Sweeney, 2013; Noble, 2018), and sentiment models and toxicity detection systems predicting negative sentiment or toxicity scores for simple statements referring to people with disabilities (Hutchinson et al., 2020).

2.3.2 Documentation as a Tool for Bias Mitigation

In response to the risks and impacts discussed in the prior sections, several groups converged on documentation as a tool to support transparency and accountability and mitigate the harms caused by ML and NLP applications. Version 1 of the data statements schema (Bender and Friedman, 2018) was released about the same time as datasheets for datasets (Gebru et al., 2018, 2021) and Dataset Nutrition Labels (Holland et al., 2018; Chmielinski et al., 2022). These documentation formats focus on the curation rationale, collection methods, and contents of datasets. The following year, model cards for model reporting (Mitchell et al., 2019), Nutrition Labels for Data and Models (Stoyanovich and Howe, 2019), and FactSheets (Hind et al., 2018; Arnold et al., 2019) were published. These formats direct attention to the algorithmic aspects of a machine learning model or product, which includes details about the data used to train the model. Many of these formats have since been honed and updated to better address their intended audiences and goals (see §3.2.1). Still further documentation formats have been proposed for specific audiences and use cases (Shimorina and Belz, 2022; McMillan-Major et al., 2021; Pushkarna et al., 2022; Hutchinson et al., 2021). Data statements stand out as a dataset documentation format specifically designed for language data types.

By the end of 2019, Bender and Friedman (2018) had been cited 26 times. At the time of writing of this dissertation, that number now sits at 627. The subjects of the early data statements include resources that had been digitized for Bornholmsk (Derczynski and Kjeldsen, 2019), a language spoken on an island in the Baltic Sea, and a dataset of Reddit posts in Danish containing conversations on rumors (Lillie et al., 2019). Other publications referencing data statements presented new documentation standards and formats (Seifert et al., 2019; Raji and Yang, 2019; Blasch et al., 2019) and efforts towards fairness, accounting for bias, and linguistic representation in ML and NLP (Mehrabi et al., 2021; Holstein et al., 2019; Selbst et al., 2019). Data statements have also been used in dataset cataloging efforts to explore the gaps in existing

data collections (Vidgen and Derczynski, 2020). The increasing visibility of documentation as a tool for accountability has given rise to retrospective analyses of existing datasets has that were developed and released by others without thorough documentation (Bandy and Vincent, 2021; Kreutzer et al., 2022; Birhane and Prabhu, 2021; Dodge et al., 2021). This use case is critical as datasets have become incomprehensibly large and infeasible for individuals to analyze by hand (Bender, Gebru et al., 2021b). These works, though not always rooted in NLP applications, further support the movement towards standardized documentation as a mitigation strategy for the negative social impacts of technology by building awareness and encouraging uptake.

Documentation is one of many tools to support developers' awareness of ethical issues in technology (Boyd, 2021). Paullada et al. (2021) note that standardization is important for rigorous documentation practices, and with this proliferation of formats, developers need guidance on how to choose the appropriate format for their project. However, imposing standardization without providing time and support for practitioners to develop their skills to meet the new standards risks adding to the barriers to participation and visibility in the field, especially for students, early career practitioners, and practitioners from regions that have fewer resources for professional development. Bender and Friedman (2018) acknowledged these potential concerns around standardization, and suggested workshops for developing training, supporting materials, and best practices for documentation.

2.3.3 Summary of Ethical Considerations in Natural Language Processing and Machine Learning

Despite academic practices that isolate technologists from other fields, a growing number of ML and NLP practitioners are calling for interdisciplinary collaborations to understand data and its ethical considerations in context. Within the fields of NLP and ML researchers are actively working to integrate engagement with ethical considerations into professional practices and to understand the negative impacts of language technology (§2.3.1). Documentation aligns with these goals as a tool to both understand the potential impacts of documented technology and support ethical professional practices (§2.3.2). As documentation formats for NLP and ML are still relatively new, further refinement of these formats requires learning from actual use cases with a variety of stakeholders. Value sensitive design, introduced in the next section, offers method-

ologies for identifying and engaging with stakeholders, analyzing qualitative data, and interpreting results, in this case towards more refined documentation formats.

2.4 Value Sensitive Design

Value sensitive design encompasses philosophical and analytical methods for examining technology, giving insight into the ways in which humanity and technology interact and influence one another (Friedman et al., 2006b; Friedman and Hendry, 2019). The foundation of value sensitive design is its tripartite methodology, which consists of conceptual, empirical, and technical investigations. Conceptual and empirical investigations produce insights, through theoretical methods and social science methodologies respectively, into the values of a technology or the values of the groups of people who create and use the technology. Technical investigations then analyze the technology itself within a particular human context.

These investigations are iterative and integrative, in that they may be used in sequence with one another in the development of a technology. For example, Friedman et al. (2000) conducted a conceptual investigation into the ways web browsers impact informed consent processes online. Millett et al. (2001) then built on Friedman et al. (2000)'s conceptual investigation in a retrospective technical investigation where they analyzed design changes in two web browsers with respect to how the changes affected web browser support for online informed consent. From the shortcomings Millett et al. (2001) identified, Friedman et al. (2002) developed a prototype tool for managing cookies in line with the values for informed consent identified in Friedman et al. (2000) and empirically evaluated the prototype in a user study. This interweaving of the different kinds of investigations exemplifies the iterative process of value sensitive sensitive design.

An approach to design that is sensitive to values must have a working definition of values and must have a method to determining whose values will be considered. In §2.4.1 I present value sensitive design's working definition of values. This definition is accompanied by examples of the possible interactions between considered values and one method for surfacing value interactions, called value scenarios. I then provide an overview of various theories and approaches to determining whose values to prioritize in technological development in §2.4.2. These approaches categorize groups of people in terms of stakeholders; I discuss the history of the term stakeholder in indigenous North American contexts as an important consideration for collaborations working with communities in the United States and Canada.

2.4.1 Values and Value Tensions

Technologies are designed with influence from particular people and their values, or “what is important to people in their lives, with a focus on ethics and morality," and the technologies then reflect those values back into the world as the technologies are used (Friedman and Hendry, 2019, pg. 24). These values are interrelated and technology designed to support one value may support many others while simultaneously being in opposition with other values. In considering the renovation of a historic building, Mok and Hyysalo (2018) used value sensitive design methodologies to investigate how a new solar energy system may be integrated into the architecture of the building. Among the values they identified, they found the tension between two, cultural heritage preservation and ecological modernization, to be of great concern to the stakeholders they engaged with. While installing the system would support the stakeholders’ goals towards energy self-sufficiency, the installation process could also cause damage to the building’s historic value or alter the building’s unique silhouette. They proposed subtle visibility in the design of the solar energy system as a way to minimize the impact of the system on the building’s appearance while still producing meaningful amounts of energy.

Identifying value tensions can lead to ideas for design innovations and future investigations. Value scenarios are a technique for imagining possible outcomes for the use of a technology and its impact on stakeholders, such as the implications of use over a long period of time or of widespread uptake of the technology (Nathan et al., 2007; Czeskis et al., 2010). Bender and Friedman (2018) use value scenarios to surface values and value tensions in the initial proposal of the data statements version 1 schema. In one of their value scenarios, the standardization and uptake of data statements at academic conferences results in fewer papers from under-represented regions and fewer papers publishing new datasets due to authors missing data statements with their publications or choosing to copy the data statements of existing datasets rather than build new datasets. This scenario highlights the tensions between inclusion and standardization; while standardization helps to fully realize the potential of data statements in supporting accountability and reproducibility, academics from regions where training in the standard is less available suffer further barriers to engaging with the broader academic field. In response to the tensions identified in this value scenario, Bender and Friedman (2018) suggest strategies for mitigating the negative impacts of the widespread uptake of data statements, including investigations into the use of data statements and mentoring programs for

building familiarity with writing data statements. Chapter 3 follows up on these suggestions by reporting on a workshop designed for NLP practitioners to both draft and exchange peer feedback on data statements and provide their own feedback on the data statements schema.

2.4.2 Stakeholder Conceptualization and Analysis

Stakeholder analyses have long been utilized as part of the value sensitive design methodology toolkit in part to identify whose values to consider in technological design and how to balance the resulting tensions (Friedman et al., 2006a, 2017). Freeman (1984) attributes the first use of *stakeholder* to the Stanford Research Institute in 1963 where it was used as a broader term than *stockholder*. It has since been used in the fields of organizational management and business ethics as a way of identifying groups and individuals who can affect or are affected by the achievement of an organization's objectives (Freeman, 1984). In the context of value sensitive design, stakeholders have come to be more broadly interpreted in context of technology instead of business organizations and may include more abstracted notions of affected parties such as societies, future generations, and natural resources (Friedman and Hendry, 2019). In considering indigenous concepts of stakeholders, Leonard (2021) also includes ancestors and artifacts as stakeholders to consider in indigenous language reclamation. Like Freeman (1984)'s definition, value sensitive design distinguishes between two categories of stakeholders: *direct stakeholders* who directly interact with the technology under investigation and *indirect stakeholders* who are affected by others' use of the technology regardless of whether they use the technology or not (Friedman and Hendry, 2019).

Stakeholder analyses are frequently conducted in conceptual investigations to define the scope of a project. Value sensitive design has analyzed stakeholders in terms of roles, where one individual may have multiple roles with respect to a particular technology, and in terms of their interaction with the technology: direct stakeholders use the technology and design decisions are made with those individuals in mind, whereas indirect stakeholders are affected by the use of the technology, but may not use it themselves (Nathan et al., 2008). Suresh et al. (2021) add to established practices of determining viewpoints to consider in technical design work by moving away from describing stakeholders by their roles in the ecosystem. They instead characterize stakeholders in terms of the type of knowledge the stakeholder brings to the context (formal, instrumental, and personal), as well as the localization of that knowledge in the context. Suresh

et al. (2021) present machine learning (ML), the data domain, and the general milieu as the possible contexts, though I believe ML may be expanded to more broadly encompass technical expertise in the processes of dataset creation without restricting the context of use to ML applications. They also provide a process for mapping stakeholders' goals with respect to the technology to considerations for design decisions. This framework makes efforts towards a more equitable treatment of the stakeholders than previous frameworks based on technical expertise alone, while also providing more visibility into the multifaceted knowledges that stakeholders bring to the design process than role-based frameworks.

In developing their Data Cards framework, Pushkarna et al. (2022) provide an alternative typology of stakeholders consisting of producers, agents, and users, arguing that the scope of Suresh et al.'s framework is too broad. Pushkarna et al. define producers as those who create datasets and dataset documentation, agents as those who may themselves interact with the datasets and documentation or who could determine how others might use them, and users as those who interact with the downstream systems built using the datasets. The authors suggest Data Cards producers should be focused on conveying information to readers who have access to ML technical expertise and recommend other documentation artifacts for readers without ML technical expertise. While they include subject matter expects and domain experts as example producers within their framework, highlighting those with access to ML technical expertise as the intended audience ignores the other kinds of expertise that may be necessary for understanding a dataset, like linguistic or cultural knowledge for a given community.

Young et al. (2019) propose another evolution of stakeholder conceptualization and engagement for the development of tech policy. In their Diverse Voices methodology, panels of experiential experts are formed to provide feedback on drafted policy documents. The Diverse Voices How To Guide describes three perspectives from which experiential experts may draw their experience: 1) lived experience experts whose expertise is based on their ability to communicate their situated experience, 2) institutionally affiliated experts who work in organizations directly supporting stakeholders, and 3) social support experts who have a personal relationship to a person with lived experience (Magassa et al., 2017). The goal of these panels is to create more inclusive policy by soliciting comments on how the policy may impact groups who have been underrepresented in policy considerations. Importantly, these panels take place prior to the document being delivered to policymakers, thereby reducing the potential for injustices and adverse impacts.

Terminological note While I use the term *stakeholder* throughout this thesis, I acknowledge that the term is not appropriate when collaborating with particular communities. Numerous indigenous communities in the Americas have argued that the term obfuscates their status as rights holders in terms of the rights to natural resources, cultural self-determination, and linguistic freedom guaranteed by local governments as well as the UN Declaration on the Rights of Indigenous Peoples (UNDRIP) (Office of the High Commissioner for Human Rights, 2007), and they do not identify in such a way that centralizes colonial organizations (Whiteman, 2009; Bruijn and Whiteman, 2010).[10] Additionally, the term was used in colonial settler practices to refer to settlers in Canada who had claimed plots of land using wooden stakes prior to treaties being negotiated with indigenous peoples.[11] A full critical analysis of the term is not within the scope of this project, so I omit it from the contents of C3DAR and leave considerations of alternative terms for future work.

2.4.3 Summary of Value Sensitive Design

In this section, I introduced value sensitive design and its tripartite methodology. Design choices in support of specific values may lead to tensions; value scenarios are an approach to identify these tensions and imagine possible paths forward (§2.4.1). Deliberately developing for the values of specific stakeholders may also help to decide design priorities when a tension is encountered. Various stakeholder analysis methods exist for identifying different stakeholders of a technical system (§2.4.2). Stakeholder is a common term in technical fields, however it has a negative connotation for some indigenous communities as a reminder of colonial settler practices and indigenous dispossession and alternative terms should be used in those contexts.

Bender and Friedman (2018) leveraged value sensitive design's methodological framework in the initial formulation of data statements. Bender and Friedman conducted conceptual investigations and technical investigations to develop the first version of data statements. Chapter 3 details the empirical and technical investigations that I undertook with Dr. Bender and Dr. Friedman towards the development of the second version of the data statements schema. More details about the methods previously used in value sensitive design

[10] See also discussions from Indigenous Corporate Training Inc. at `https://www.ictinc.ca/blog/9-terms-to-avoid-in-communications-with-indigenous-peoples`.

[11] According to the government of British Columbia's Writing Guide for Indigenous Content at `https://www2.gov.bc.ca/gov/content/governments/services-for-government/service-experience-digital-delivery/web-content-development-guides/web-style-guide/writing-guide-for-indigenous-content/terminology`.

research and in the development of data statements are provided in §3.2.3. In §3.7 I present a co-authored discussion of the ways in which the empirical and technical investigations toward the data statements Version 2 schema contribute to methods for co-evolving technology and technical community practice. Chapters 5 and 6 report on a retrospective technical investigation and a technical investigation, respectively, towards the development of C3DAR.

2.5 Summary

In this chapter I summarized the various literatures that inform this dissertation and present introductions to indigenous research, signed language research, ethical considerations in NLP and ML, and value sensitive design. In §2.1 I provided a brief history the complex relationship indigenous peoples have with research conducted by researchers and groups outside of the community, particularly indigenous peoples in the United States. This history begins with the hegemonic research that was conducted on indigenous communities that objectified them and facilitated their subjectification to colonial powers and genocide (§2.1.1). In the United States, ongoing federal efforts toward the revitalization of indigenous languages and cultural practices have produced mixed results, though recent legislation offers more hope in providing funding for indigenous-led language centers (§2.1.2). Language research has also contributed to exploitative practices against indigenous communities; linguists who are themselves indigenous and linguists who have long term working relationships with indigenous communities have made clear the harms of these practices and the benefits of more collaborative models of research, changing perceptions in the field about the positions of researchers with respect to community members (§2.1.3). Indigenous people are increasingly leading research and policy development following their own methods and values and in support of their own community needs using frameworks such as IDS (§2.1.4). In §2.1.5, I considered the implications of the colonial history of language research and the current topics voiced by indigenous community members on future directions for indigenous language technology development and IDS.

Parallels between indigenous communities and signed language communities have been drawn in terms of their quests for linguistic rights (§2.2). English and other majority spoken languages have been systematically imposed on deaf communities in education and through medical interventions, despite signed languages being more accessible to deaf people than spoken language (§2.2.1). Connections between signed

language studies and indigenous studies are surfaced in §2.2.2, though each still have their own unique challenges. The challenges for signed languages include legal recognition (§2.2.3); competing models and perceptions of deafness have resulted in various kinds of legislation around the world. In §2.2.4, I summarized current sign language processing technologies and ways they could be improved. For future signed language projects, I presented open questions for the intersection of ethical practices for signed language research and technology (§2.2.5).

The ethical considerations for community-specific development are informed by the ongoing conversations technical fields are having about broader ethical topics relating to language technologies (§2.3). I presented the harms of language technology and the frameworks for understanding them in context in §2.3.1. One proposed mitigation strategy for these harms, documentation, has started to gain traction (§2.3.2), but investigations into more use cases and the needs of various stakeholder groups are needed before documentation standards can be implemented.

One such documentation tool, data statements, was developed using the methodologies from value sensitive design. Value sensitive design aims to shape technology and community practice iteratively and with attention paid to the values held by various stakeholder groups (§2.4). Value scenarios may be leveraged to imagine interactions between different values and support design decisions that balance values that have been identified as key priorities for technical development (§2.4.1). Stakeholder analyses are one method for determining these priorities (§2.4.2). Various methods support categorizing potential groups who may use or be indirectly impacted by the use of the technology being developed.

The work presented in the following chapters continues the evolution of documentation, collaboration, and community practice. Building on Bender and Friedman (2018), Chapter 3 introduces an investigation I conducted with Dr. Bender and Dr. Friedman into the needs of NLP practitioners with regards to language dataset documentation and how to support those needs with best practices, resulting in the development of Version 2 of data statements for NLP. Chapter 4 then presents how I developed a documentation format expanding on data statements Version 2 to support collaboration between NLP practitioners and indigenous and signed language communities.

Chapter 3

Data Statements Version 2

This chapter presents work done collaboratively with Dr. Emily M. Bender and Dr. Batya Friedman (included in this dissertation with their permission). It is a modified version of the article we published in the Association for Computing Machinery (ACM) Journal on Responsible Computing:

> Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data Statements: From Technical Concept to Community Practice. *ACM J. Responsib. Comput*. 00, JA, Article 00 (May 2023), 17 pages. https://doi.org/10.1145/3594737

Amid several documentation toolkit proposals, data statements for NLP (Bender and Friedman, 2018) began to see uptake within the NLP community in the year following its proposal in response to calls for responsible computing practices (see Chapter 2). Seeking to encourage further uptake and to hone the proposed toolkit to better address the practical needs of the NLP community, we engaged the community in two-way knowledge sharing: improving the technology based on community insight while training community members in its use. We explored how to improve the toolkit in three senses: (1) the content of the toolkit itself, (2) engagement with professional practice, and (3) moving from a conceptual proposal to a tested schema that the intended community of use may readily adopt. To achieve these goals, we first conducted a workshop with NLP practitioners in order to identify gaps and limitations of the toolkit as well as to develop best practices for writing data statements, yielding an interim improved toolkit. Then we conducted an analytic comparison between the interim toolkit and another documentation toolkit, datasheets for datasets. Based on these two integrated processes, we presented our revised Version 2 schema and best practices in a guide

for writing data statements. Our findings more generally provide integrated processes for co-evolving both technology and practice to address ethical concerns within situated technical communities. This Version 2 schema is the foundation from which C3DAR is developed in the later chapters of this dissertation.

3.1 Introduction

As discussed in §2.3, data statements were initially proposed by Bender and Friedman (2018) in response to growing awareness of the fact that machine learning (ML) approaches to language technology bring various risks of harm to both system users and others affected by system use (Sweeney, 2013; Caliskan et al., 2017; Noble, 2018). ML approaches to any problem involving data created by or about humans have similar risks, but both the particular risks and the ways in which they connect to data collection practices differ by data type. Data statements, which are honed to their data type, are part of a wave of convergent dataset and model documentation proposals (see §3.2) that seek to position technologists, those who procure and deploy technology, and community members to mitigate potential harms by providing transparency into the data used for training and testing such systems.[1]

Toolkits refer to physical and digital materials that support people in carrying out methods and processes (Hendry et al., 2021). Considering documentation toolkits and their purposes, no toolkit can produce any benefit if people don't use it to create documentation. Furthermore, the benefit of the documentation will be limited if it is not sufficiently detailed, nor accessible. So we asked two intertwined questions: How do we adapt our proposed toolkit and practice so that it is feasible for the practitioners we hope will take it up to do so? And: How do we facilitate community uptake? In this paper we present both the ways in which we engaged and learned from the community and the resulting improved toolkit, including a revised schema and distilled best practices. We present the revised schema and best practices in a guide which we developed to support data statement authors in creating documentation accessible both to technologists and to third parties who need or want to understand data used to construct technology. The schema and best practices are available in Appendix A.[2] As is evident from our characterization of these outcomes, we view the toolkit and associated practices as a single intertwined system. Most broadly, our contributions speak to how to

[1]Data statements also support the use and understanding of NLP systems built and evaluated with small datasets in resource-constrained scenarios, where ML may not be applicable.

[2]The guide (Bender et al., 2021a) is also available at `http://techpolicylab.uw.edu/data-statements/`

evolve both technology and practice to address ethical concerns within situated technical communities.

We structure the paper as follows: In §3.2 we present an overview of recent documentation proposals for datasets, models, and systems and situate data statements within this ecosystem, as well as a review of the methodologies that we draw from value sensitive design (Friedman and Hendry, 2019). We lay out our researcher stance, research questions and specific methods in §3.3, §3.4 and §3.5, respectively. §3.6 gives an overview of the revisions to the toolkit and in §3.7 we provide reflections on both our methodology and what we learned about how data statements fit into the landscape of data documentation practice and into practitioners' activities. Finally, in §3.8, we provide an outlook onto future work, including engaging with a broader set of stakeholders, further study of the uptake and use of data statements and generalizations to other data types.

3.2 Background

3.2.1 Documentation Toolkits

In response to a wide range of potential harms from applying pattern recognition ("AI") at scale, in 2017-2019 several research groups, mostly from the United States by affiliation, began to develop documentation toolkits to support transparency in AI systems. As shown in Table 3.1, each of these documentation toolkits was developed with inspiration from a particular non-digital documentation format and with particular users, harms and use cases in mind.

More recently, as documentation toolkits gain traction, we are seeing two trends. First, documentation toolkits are being integrated into standard practice and early-stage standards to mitigate and manage bias in AI systems (Schwartz et al., 2022). Second, initial documentation toolkits are being revised as part of iterative design processes, leading to more formalized and complete versions. For example, based on feedback from legal scholars and user studies, the categories and questions employed in datasheets have been refined (Gebru et al., 2021); the Data Nutrition Project updated their Data Nutrition Label tool to include intended use cases (Chmielinski et al., 2022); and IBM expanded their FactSheet to include specialized template development for project teams (Richards et al., 2020). In this second stage of documentation toolkit development, the field is moving beyond initial toolkit formulation to explore the needs of documentation

Toolkit	Inspiration	Focus	Ref
Datasheets for Datasets	Electronics documentation for components, etc.	Datasets: detailed documentation on key dataset design issues; intended for experts	Gebru et al. (2018, 2021)
Data Nutrition Project	Standardized nutrition labels for prepared food	Datasets: brief standardized format for details on the construction and contents of a dataset; intended for experts and non-experts	Holland et al. (2018), Chmielinski et al. (2022)
Data Statements for NLP	Description of participants in social and medical research	Datasets: highlights the design, the people represented, and considerations that arise from use of language data types	Bender and Friedman (2018)
Nutrition Labels for Data and Models	Standardized nutrition labels for prepared food	Datasets and models: automatically calculated information about data and models to inform on production processes behind ML models	Stoyanovich and Howe (2019)
Model Cards for Model Reporting	TRIPOD statement proposal in medicine	ML Models: model characteristics including type, use case, performance variance and performance measures; complement to datasheets	Mitchell et al. (2019)
FactSheets	Suppliers Declaration of Conformity (e.g. telecom, transportation)	AI model or service: Purpose and criticality of a model; measures of a dataset, model or service; creation and deployment process	Arnold et al. (2019)

Table 3.1: Documentation Toolkits: Inspiration and Focus

writers, including addressing gaps and lack of clarity in the initial toolkit directions and support for skill development in writing, reading and using the toolkits. The work reported on here contributes to these second stage efforts.

3.2.2 Data Statements

A data statement consists of schema elements and is defined by Bender and Friedman as "a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software" (Bender and Friedman, 2018, p.587). The Version 1 schema consists of two parts: a long form and a short form. The long form contains nine schema elements, which each correspond to a set of questions or suggested descriptions about an aspect of the dataset, such as the curation rationale, language variety or demographics of the speakers in the dataset (Bender and Friedman, 2018, §5). The short form is a summary of the long form designed to be used in publications that reference the dataset. Practitioners are encouraged to use both forms in coordination with papers introducing datasets, as part of reports of experiments that used a dataset, and alongside documentation for a model trained on a dataset. Data statements have been used in dataset cataloging efforts to explore the gaps in existing data collections (Vidgen and Derczynski, 2020), and recent work with datasheets points to documentation's ability to support developers' awareness of ethical issues in ML technology (Boyd, 2021). For an illustration of both Version 1 and Version 2 of the data statements schema as well as a sense of how they differ, see Figure 3.1 on page 61, discussed further in §3.6.

3.2.3 Value Sensitive Design

Value sensitive design is an established approach for foregrounding human values and well-being in the technical design process (Friedman and Hendry, 2019). Value sensitive design takes a broad stance in defining technology as a combination of tools, technologies, and infrastructure that shape human activity, encompassing both physical and digital artifacts (Friedman and Hendry, 2019). Mok and Hyysalo (2018) used value sensitive design methodologies to integrate a new solar energy system into the architecture of a historic building, while Millett et al. (2001) employed value sensitive design methodologies to improve

informed consent features in internet browsers. At the core of value sensitive design is the tripartite methodology of iterative and integrative conceptual, technical and empirical investigations, as well as the practical strategy of co-evolving technology and social structure (including community practice). This methodology allows for extended inquiry into the interaction between technology and society through iterative investigation and evaluation over time. For example, the retrospective analyses that Millett et al. (2001) conducted were built on the conceptual investigation described in Friedman et al. (2000) and themselves were the foundation for technical interventions that were empirically evaluated in Friedman et al. (2002). Similarly, two of the authors of this paper, Bender and Friedman, employed this approach in the initial development of data statements Bender and Friedman (2018). We continue to draw on value sensitive design for the subsequent work presented here.

In their initial work, Bender and Friedman began with a conceptual investigation, drawing on the definition of bias presented in Friedman and Nissenbaum (1996) as "systematic" and "unfair discrimination." They paid particular attention to how bias in computing systems could reflect preexisting social conditions or emerge over time when computing systems developed for a specific set of circumstances and populations were used in other circumstances and with other populations. As a proof-of-concept and technical investigation, Bender and Friedman then applied the data statements toolkit to two actual datasets, one of English Twitter data and one of English and French video interview data. In addition, they employed value scenarios (Nathan et al., 2007; Friedman and Hendry, 2019) as a conceptual method to explore how an at-the-time imagined documentation toolkit could provide benefit both in terms of mitigating bias and contributing to better science. Value scenarios provided a structured way of envisioning futures, bringing forward both potential positive and negative impacts of a not-yet-built-and-deployed technology on individuals, communities, fields and societies. One of their value scenarios, concerning the potential for data statements to become a force for exclusion if standardized too quickly, led Bender and Friedman to call for empirical investigations exploring how data statements as a practice would work for a diverse range of practitioners.

In the work reported here, we follow up on this call. In doing so we leaned further into value sensitive design's tripartite methodology. With the goal of improving the 2018 data statement schema from a community-of-use perspective, we first conducted an empirical investigation with one direct stakeholder

group,[3] NLP dataset creators, to gather their perspectives and insights for how data statements and the surrounding practice could be improved by clarifying existing schema elements, identifying gaps where additional schema elements were needed and collecting best practices. Our empirical work was followed by two sequential technical investigations to revise the data statement schema. In the first we used empirical workshop results to guide reformulation of the schema and identification of best practices; in the second, we compared datasheets for datasets to the reformulated schema to identify and fill any additional gaps.

3.3 Researcher Stance

Our research team based in the United States is comprised of computational linguists facile with NLP and ML systems and an information scientist skilled in the application of value sensitive design, particularly around mitigating bias in computing systems. All team members previously participated in developing documentation toolkits for datasets used in ML systems.

3.4 Research Questions

In moving from an envisioned documentation toolkit to one positioned to be taken up by a research community, we sought to make the data statements toolkit more robust with respect to institutional contexts, researcher backgrounds and research goals. This motivated two broad research questions:

1. How should the data statements for NLP schema be updated to better support the range of projects it might be used for in the international NLP community?

2. How could we support practitioners in a wide range of institutional contexts in writing data statements and facilitate community uptake of this practice?

[3]Here we distinguish between *direct* stakeholders who interact directly with the documentation toolkit either by writing or reading documentation and *indirect* stakeholders who may never see the resulting documentation but nonetheless are affected by others' use of it (Friedman and Hendry, 2019).

3.5 Methods

To gain traction on these research questions, we took a two-phased approach, drawing on a similar methodological approach from Friedman et al. (2006c). In Phase 1, to understand how NLP dataset creators would make sense of and utilize the existing schema (Version 1) we organized an empirical investigation in the form of an international community-based workshop with NLP practitioners (described in §3.5.1). Based on the Phase 1 workshop results, we developed an interim revised schema in a technical investigation. Then in Phase 2, to learn from others' efforts developing documentation proposals, we conducted a second technical investigation in which we carried out a close, analytical comparison between the schema and a related documentation toolkit (§3.5.2). Throughout, we paid particular attention to (1) how NLP dataset creators could effectively collect the information required for data statements; (2) identifying and developing heuristics for writing data statements; (3) managing privacy and ethical considerations, particularly those tied to small or vulnerable populations; (4) how data statements relate to other existing practices in the NLP community; and (5) how to document legacy datasets.

3.5.1 Phase 1: NLP community-based workshop

To uncover the strengths, gaps, confusions and limitations of the Version 1 schema elements (as published in Bender and Friedman, 2018) as well as to generate best practices for writing data statements, we held an international workshop with members of the NLP community. The workshop was accepted as part of the 12th Language Resources and Evaluation Conference (LREC); due to COVID-19 and the eventual cancellation of the conference, the workshop was held virtually over three days, May 11-13, 2020. In this empirical investigation, we sought feedback from NLP dataset developers in order to evaluate the data statement schema in practice.

Participants and their datasets We recruited participants through an open invitation over standard workshop announcement channels for the NLP community. Specifically, we invited NLP community members to a working meeting where they would engage in writing data statements. We recruited as broadly as NLP workshop distribution channels would allow, in the hopes of getting a very broad range of perspectives, and succeeded in attracting participants from around the globe, though some regions (Europe, the US) were

more represented than others. In total, 38 practitioners from 16 countries participated, including practitioners from Argentina, Mauritius, Sri Lanka as well as the US and Europe. Half (50%) of the participants identified as senior researchers, while 36.8% identified as junior researchers and 13.2% did not provide a response. The workshop was designed around training language technology practitioners. Though we had one participant who came from a different research community (legal scholarship), for the most part, there was considerable shared common ground in the academic training of our participants. This both facilitated productive working sessions and shaped the range of ideas elucidated in those sessions.

Most participants brought datasets to document; where multiple participants represented the same dataset, we considered them part of the same participant team. In total, there were 29 datasets, reflecting the collective geographical diversity both in terms of the language and content of interest. Just over half of the datasets were collections of varieties of English; other languages represented include Arabic (a mix of Arabic language varieties), Argentinian Spanish, Basque, Javanese and Yoruba, to name a few. The genre of data ranged from Twitter posts to biomedical data to proverbs.

Workshop structure and procedures The design of the workshop was driven both by our goal of eliciting formative feedback on the data statements schema as well as our goals of providing a useful training and networking experience for the workshop participants. It was also shaped by the fact that it took place over Zoom, early in the global experience of the COVID-19 pandemic. In this context, we sought to balance in-depth paired participant interactions with larger group work. We intended for participants to experience the process of writing and evaluating data statements within a peer review process, and then reflect upon and discuss those experiences with others. Towards the goal of providing a networking opportunity for participants across this international community, we designed workshop activities that we expected to provide opportunities for relationships to form, assigning new participant pairings over the course of the workshop.

The virtual workshop met synchronously in Zoom for six hours total, in 2-hour sessions across three contiguous days. In addition to these synchronous meetings participants completed some work asynchronously between sessions, as preparation for the next meeting. On Day 1, participant teams were introduced to each other and informed of the workshop's twofold goals: (1) for each participant team that brought a dataset to leave the workshop with a solid, if not complete, draft of a data statement for their dataset; and (2) for the workshop participants as a whole to identify improvements to the Version 1 schema elements and generate

best practices for writing data statements.

To achieve these ends, we formed small groups of participants around the datasets they brought, with 1–2 datasets per group. In addition, the data statement construction process was supported with a shared digital worksheet presenting the Version 1 schema elements. For each element, the worksheet provided the element explanation (from data statements Version 1, as specified in Bender and Friedman, 2018) and allowed for (a) notes; (b) draft text; (c) feedback; and (d) advice for future data statement authors.

The workshop flow was as follows. On Day 1, after the introductions, we put the participant teams into small groups to develop the first four schema elements using the worksheets. During this writing process, participants took on one of two roles: data statement "author" or "interviewer". The data statement author role entailed writing the actual schema elements for a particular dataset. The interviewer role entailed asking the data statement author questions about the dataset, to bring forward aspects which might need clarification, greater specification or were deemed unnecessary or redundant. In this sense, the schema elements functioned as questions to be asked by the interviewer and answered by the data statement author. Notes from this interview process were recorded on the worksheet. As "homework", participants finished drafting these schema elements. On Day 2, participants worked in small groups to review the schema elements drafted the day before and then in a second small group session repeated the drafting process for the remaining five schema elements, again finishing the drafting as homework. On Day 3, a final small group session allowed for peer review of the second set of elements. Finally, four breakout groups comprised of 8-9 participants with one facilitator met to reflect on the specific workshop activities and on data statements more generally. In these groups, participants were asked about topics such as what advice they would give to future data statements writers, what improvements they would like to see to the schema elements, potential uses as well as harms and misuses of data statements, and suggested best practices. Participants were therefore asked for their suggested best practices having just experienced the process of iteratively improving their own data statements and also providing feedback on others' drafts.

The materials from the workshop that served as the empirical basis for our analysis included recordings of the final breakout sessions and the short full-group debriefing sessions at the end of Days 1 and 2, as well as the data statements produced by the participants, the notes they included in their worksheets, and the notes they provided in the discussion questions worksheet for the breakout sessions on Day 3. We did not

create Zoom recordings of the small-group work on actual data statement development, as we believed that might have been perceived as intrusive and counter to the goal of building relationships among participants.

Data analysis Using an inductive process (Corbin and Strauss, 2008), we systematically reviewed the recorded material on participants' worksheets and the group discussion transcripts to identify and consolidate potential improvements to the schema and best practices. Specifically, two members of our research team with deep knowledge of language datatypes and NLP systems annotated the worksheets for tips and suggestions as well as for strengths and weaknesses in the participant-written data statements, paying particular attention to where difficulties occurred as a result of the schema definitions and scope. The lens that we used to examine the participant-written data statements was how well and completely they addressed the schema element questions, with an eye toward potential sources of bias. We also attended to overshoot: material that went beyond describing the dataset itself to include background information which would be better placed elsewhere. In evaluating the strengths and weaknesses of participant-written data statements, we found patterns that led us to develop best practices (either as practiced by data statement authors or that would have helped data statement authors). We also observed instances in which the Version 1 schema was ill-suited to certain kinds of language data, as in the case of translation data where participants needed to describe the characteristics of two (or more) languages. For the group discussion transcripts we annotated ideas around best practices. We excluded as out of scope participant comments about creating datasets (rather than documenting them) and automatic generation of data statements. Based on the analysis of these two data sources, we revised the Version 1 schema and we created general and element-specific best practices.

Interim products The data analyses and subsequent revisions resulted in the Version 2 (Phase 1) data statement schema and a draft guide for writing data statements (see §3.6 for details).

3.5.2 Phase 2: Analytical Comparison to Datasheets for Datasets

To check for completeness and make the data statement schema and best practices from Phase 1 even more robust, we followed a strategy of leveraging a related model (Friedman et al., 2006c), in this instance another documentation toolkit effort. In choosing a documentation toolkit for comparison, we sought one that also engaged with datasets (as opposed to other aspects of systems) in a detailed manner and, ideally, from

another organizational and/or institutional context as a means to enrich our development work thus far. Of the documentation toolkits described in §3.2.1, datasheets for datasets (Gebru et al., 2021) (we used v7 of the paper on arXiv, i.e., Gebru et al., 2020) is the most similar to data statements. As shown in Table 3.1, only two others pertained solely to datasets. Of these, data nutrition labels were designed to be 'at a glance', where datasheets provided more detail and thus made a better point of comparison to data statements. Datasheets were developed by industry researchers within a large tech company rather than in the academic research community, so we expected that they would capture different contextual and organizational perspectives, aligning with our stated research questions. Datasheets have also seen a high degree of uptake within the community. For example, the Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track recommended that datasets published at their venue be accompanied by documentation following datasheets for datasets, data statements, or data nutrition labels in both 2021 and 2022. This research community interest in datasheets has continued, as evidenced by the datasheets publication having over 1000 citations at the time of writing this paper.

In this technical investigation, we paid particular attention to how each toolkit conceptualizes what data is, who is writing documentation, who is reading documentation, what risks are being mitigated and what other purposes the documentation serves. To situate the comparison in the details of the two toolkits, we sought to account for each of the questions the datasheets schema asks documentation authors to consider, mapping datasheets questions to data statements elements where possible. Where there was no corresponding element in our Version 2-Phase 1 schema, we either identified a location where the information could be added to the data statements schema, or marked the question as out of scope for data statements. We found information to be out of scope for different reasons, e.g., because it doesn't pertain to language data or because we believe it would be provided in complementary documentation to data statements, such as documentation for important ethical review processes (institutional review boards (IRBs) or otherwise).

3.6 Final Products: Revised Schema, Best Practices, and Guide

The NLP community-based workshop (Phase 1) and comparison with datasheets for datasets (Phase 2) resulted in three products: (1) a revised schema (Version 2); (2) a list of key terms and best practices (general and element-specific) for writing data statements for NLP; and (3) a guide for writing data statements for

Revisions	Phase 1: Workshop	Phase 2: Datasheet Comparison
General Best Practices	New	-
Key Terms	New	-
Schema Elements		
1 Header	New	Updated c
2 Executive Summary	New	-
3 Curation Rationale	Updated b, c, d	Updated c
4 Documentation for Source Datasets	Updated a, b, c, d	Updated c
5 Language Varieties	Updated a, b, c, d	-
6 Speaker Demographic	Updated b, c, d	-
7 Annotator Demographic	Updated b, c, d	-
8 Speech Situation and Text Characteristics	Merged and updated a, b, c, d	
9 Preprocessing and Data Formatting	New	Updated c
10 Capture Quality	Updated a, b, c, d	-
11 Limitations	New	-
12 Metadata	New	Updated c
13 Disclosures and Ethical Review	New	-
14 Other	Updated b, c, d	-
15 Glossary	New	-

Table 3.2: Revisions by source of change. Each element is comprised of a: (a) title, (b) rationale, (c) description, and (d) best practices. "New" refers to the addition of an entirely new element.

NLP that presents (1) and (2) in a cogent manner. As shown in Table 3.2, the vast majority of the revisions were the result of the community-based workshop.

Revised schema (Version 2) The community-based workshop (Phase 1) resulted in the creation of 7 new schema elements as well as updates to the rationale, description and best practices of the other original 9 schema elements. In addition, the schema elements were reordered and reorganized; in one instance two elements were merged into one, resulting in a total of 15 schema elements in Version 2. These changes emerged from both explicit comments and feedback from the workshop participants as well as our data analysis. For example, 5 of the new schema elements (Preprocessing and Data Formatting, Limitations, Metadata, Disclosures and Ethical Review, and Glossary) were suggested by participants during the group discussions. Our analysis of the participants' data statements resulted in the additional Header and Executive Summary schema elements, as well as merging the Speech Situation and Text Characteristics schema elements into one element. The comparison with datasheets for datasets (Phase 2) yielded five additional revisions; all of these were to element descriptions. To illustrate the substance and depth of changes from

Version 1 to Version 2, we present the changes made to two of the schema elements: Curation Rationale and Recording/Capture Quality.

The top part of Figure 3.1 shows the changes we made to the Curation Rationale schema element. (1) *Element order*. As it was the first element in the Version 1 schema, we observed that workshop participants tended to overload the element with introductory information about the dataset. In response, we made the Curation Rationale the third element, after the new Header and Executive Summary schema elements that allow for more context about the contents of the dataset. (2) *Motivation*. Originally, motivation for how the schema element serves the reader of a data statement came after the description of the content for the element. We moved this motivation to the start of the element in the *Why* section, and included additional motivation for how the Curation Rationale also supports dataset creators. A few other schema elements in Version 1 also included motivation for why the element was included in the schema; we made this consistent across all schema elements in Version 2, including a rationale for both writers and readers in the *Why* section for each element. (3) *Elaboration*. Finally, we drew from the analyses of both phases to add more clarifying questions such that a completed Curation Rationale may better support surfacing sources of societal and/or emergent bias that may be encoded in the dataset.

While the Curation Rationale retained the original conception of the element from Version 1 (with elaborations), the changes made to the Capture Quality schema element (formerly the Version 1 Recording Quality element) illustrate a considerable re-imagining of scope and, hence, name and description of the element. Figure 3.1 also shows the two changes we made to the Capture Quality schema element. (1) *Scope*. In analyzing workshop participants' data statements, we found that this element — originally designed to capture technical biases related to audiovisual equipment used — was used creatively to document a wider variety of technical considerations. These include systems used for correcting optical character recognition (OCR) output, API reliability when requesting data from online platforms and data degradation stemming from linked data becoming inaccessible. Accordingly, we broadened the element's scope to include these and other possible sources of technical bias when capturing observations of language use in the world for use as data in a dataset. (2) *Rationale*. As described above, we added the *Why* section to convey the importance of these considerations to both data statement readers and dataset creators.

Figure 3.1: Sample elements from Version 1 vs. 2 schema. Orange represents change of element order or title; green reorganization within an element; and blue elaborations to content

Best practices Our advice to data statement writers takes the form of best practices, identified through analysis of workshop participants' reflections as well as the strengths and weaknesses of the participant-written data statements produced during the event. There are 16 general best practices which are applicable across data statement elements or otherwise pertaining to the data statement as a whole. In addition, there are 47 element-specific best practices, ranging per element from one (for Speech Situation and Text Charac-

teristics, Other and Glossary) to nine (for each of Speaker and Annotator Demographics). The best practices convey three levels of emphasis, distinguished linguistically: (1) Best practices we believe must be followed to create a successful data statement, articulated as imperatives. (However, in many cases, the imperative instruction is to *consider* a course of action.) (2) Best practices we strongly advise, expressed with *should*. (3) Best practices we propose as one good way to proceed, expressed with *recommend*. The determination of which level of emphasis to use for each best practice was decided through deliberation among the three authors on what information we thought would be feasible for data statement authors to provide in most contexts as well as what information data statement readers would need to answer questions relating to possible sources of bias.

As an illustration of the best practices, here is general best practice #4, which reads:

> *Some of the data statement elements concern information that may require advanced planning to collect (e.g., demographic information). We recommend determining what information is to be collected and how at the start of the project, leaving time for ethics review board approval as appropriate.*

This best practice is derived from workshop participant comments that advocate working on the data statement early in the dataset development process, e.g., "Recommend drafting the data statement during the data creation process, as some information is more easily available at the time than later." This general best practice also reflects a proactive response to dataset creators who may feel uncomfortable about collecting and handling demographic information, even while understanding the importance of such information for creating representative datasets.

Guide for writing data statements To assist technologists, scholars and others with writing data statements using the revised schema (Version 2) and drawing on the best practices, we created a third product which took the form of "A Guide to Writing Data Statements: For Natural Language Processing" (Bender et al., 2021a). The guide brings together the Version 2 schema elements and best practices into one integrated document that is organized to support the data statement writing process. General best practices (a total of 16) that cut across all aspects of the data statement writing process appear first, followed by key terms germane to language data types: annotator, disordered speech, elicited data, found data, language data, lan-

guage variety, speaker, speech synthetic text and text. Next come the 15 schema elements, each on its own page. For each element, we provide a rationale (the *Why*), a description (the *What*) and element-specific best practices. Most pages have ample white space for note-taking and the user's annotations. A sidebar "schema-map" acts as a memory aid and facilitates flipping among related elements. The guide concludes with two appendices, the first for converting schema Version 1 to Version 2; the second for situating data statements with respect to other documentation toolkits.

3.7 Reflections on Process and Products

On methodology We took a two-phased approach, engaging first with NLP practitioners directly in the context of their own work writing data statements and then conducting a comparative analysis with a closely related documentation toolkit. In reflecting on our methodological strategy, we can make several observations. First, following value sensitive design's tripartite methodology, the two approaches we employed represented different types of investigations. Specifically, the workshop was an empirical investigation which positioned participants to directly engage with the data statement writing process and share their insights and advice in addition to the data statement artifacts they generated for their own datasets. As such, this empirical method invited participant creativity and allowed participants to express themselves in whatever ways they wished. We then followed this with two technical investigations: the Phase 1 reformulation of the schema and development of best practices and the comparative analysis with a closely related documentation toolkit. The comparative analysis focused on the technical structure and details of the two toolkits. This technical method afforded systematic and comprehensive surface-level comparison and was well positioned to shine a light on omissions in the interim Version 2 schema. Second, employing empirical and technical approaches in tandem yielded a broader set of improvements than either approach would have in isolation. Others wishing to improve similar toolkits might wish to employ a similar strategy: engaging a combination of empirical and technical investigations.

Considering the workshop further, we next call out two aspects of special interest: one following from participant make-up, the other from process. In terms of participant make-up, engaging directly with NLP practitioners from different countries and different institutional research contexts provided us with access to their collective wisdom and creativity. As a group, their depth and breadth helped us understand where the

Version 1 data statement schema could be improved to better meet a wide range of needs and backgrounds, how data statement writers could be better supported with a structured and detailed guide to writing data statements and which key insights and best practices to share with others. This was particularly valuable given our goal of creating a documentation toolkit which would be accessible to researchers from institutional contexts different from US academia. Researchers based in different cultures helped us learn about different ways in which particular kinds of data about speakers and annotators might be considered sensitive as well as different levels of institutional support around ethics review. These lessons informed both the design of the schema and the best practices we articulated. In terms of process, the practice of interviewing a dataset developer as a means to elicit meaningful documentation led us to a more general observation: namely, that interviewing by an outsider serves as an effective method for eliciting content from dataset developers at a meaningful level of granularity. Where the term "documentation" usually evokes a dry asynchronous practice — the documenter writes, later others read — we found the interview technique made data statement writing interactive and as a result more rewarding for both the data statement authors and (future) readers.

On automation Among the suggestions from our workshop participants (and other members of the NLP community who we have spoken with about data statements) were those concerning automation and data statements. There are two variants: We might ask to what extent the production of data statements can be automated and to what extent data statements might be rendered automatically processable. In both respects, we see value in keeping this process manual. For the former, we believe that writing a thorough and beneficial data statement requires engaging thoughtfully with the data being documented, whereas automation tends to produce distance between author and dataset. For the latter, it is very important that data statements remain designed to be accessible to human readers, from a wide range of stakeholder groups. Designing them for automatic processing would likely render them less readable. In this sense, we see data statements are very much complementary to other kinds of metadata, such as the Dublin Core metadata standards (Technical Committee ISO/TC 46, Information and documentation, Subcommittee SC 4, Technical interoperability, 2017, 2019). Such standards support discoverability of datasets; a data statement provides the reader who has discovered a dataset of interest with information about its *content* and *context*. That said, data statement authors are encouraged to use BCP-47 language codes which would allow for automation to

determine which languages are represented in the data catalog and, importantly, which are not yet represented. As envisioned in the original data statements paper (Bender and Friedman, 2018), that information would position the field as a collective to systematically fill gaps for underrepresented languages. Consistent with the sentiments above, this particular automated task would not interfere with the benefits of a primarily manual cataloging process.

On unanticipated use cases From our perspective, one of the more interesting outcomes concerned use cases for data statements. Recall that data statements were envisioned to mitigate the harms of exclusion and bias in language technology and support transparency in future applications of that technology through informed dataset selection, more thorough dataset analysis and bringing the ethical considerations of NLP data to the foreground for all NLP practitioners (Bender and Friedman, 2018). That said, the workshop participants identified several other use cases including functioning as an analogy to code README documents in increasing the accessibility of datasets, increasing the accessibility of NLP research to other fields, contributing to data repository metadata and serving as a planning tool for careful dataset development. These unanticipated uses point to the need for more general support of dataset development, integration and communication and increased valorization of the work that goes into data creation and dataset maintenance (Sambasivan et al., 2021).

On situatedness of documentation practices Our comparison to the datasheets schema allowed us to see some of the ways in which the initial development context of data statements shaped the resulting toolkit. Two key features of that context are that data statements (both Version 1 and Version 2) were developed from the perspective of academia and with a concrete focus on language datasets. We see the impact of the academic context in the way that data statements seek to complement rather than encompass work done by IRBs, incorporating a place for a pointer to any IRB documentation in the Disclosures and Ethical Review element.

We find that our specific focus on language data enabled several key features of our toolkit. First, we are able to provide prompts in the schema for particular kinds of information that are relevant to issues of emergent bias with language datasets (e.g., dialect, genre). Second, we have a clear distinction between data (language produced by language users) and annotations (any additional labels added to that language data),

and we prompt for information about the people involved in each process. Separating these out, we argue, will position dataset and technology users to better diagnose the source of problems as they arise. Third, and possibly most importantly, by grounding our toolkit in a specific data type, we are able to make our recommendations more concrete, which in turn makes data statements easier for dataset producers to write and for data statement readers of all backgrounds to understand.

On productive friction The work reported here is the product of an interdisciplinary team. Authors McMillan-Major and Bender are computational linguists; author Friedman is a designer and technologist with expertise on human values in technical design. Navigating our interdisciplinary discussions was difficult and time consuming. We found that it was easy to misunderstand, both at the level of vocabulary and at the level of work behind the results from the other field. However, at the same time, we found that the resulting friction was generative, and taking the time to reach understanding led both to valuable new insights and to research products accessible to broader communities. For example, we developed the key terms in the data statements schema both to aid our own mutual understandings as well as to support non-NLP experts in their engagement and work with data statements. Ultimately, we found that the interdisciplinary experience brought value even beyond meeting this necessity: attending to the turbulence rather than trying to push past it and extending grace and respect across the disciplinary differences brought us benefits in the form of learning opportunities and insights that come from having to actively work towards clarity and mutual understanding.

On standardization: why, what and when? Those differing contexts of documentation schema development, varied targeted objects for documentation and disparate experiences of the developers themselves have resulted in a proliferation of diverse documentation schemas. With all of these different formats come challenges for coherent and widespread uptake of documentation. While standardization towards a few documentation schemas offers one way forward, it raises yet another set of questions: Should the schemas themselves or just the content of the documentation be standardized? At what jurisdiction should documentation be standardized, especially within interdisciplinary fields where contexts and data types may vary greatly? In the case of NLP, language data in the form of text is often accompanied by video and image data, which carry their own unique considerations for bias and ethical data management. Is it time to converge

and standardize now, or is it better leave time for additional innovation and standardize at some point in the future? What rhythms of the innovation-convergence-uptake life cycle should we consider, which should we avoid? While institutions involved with standardization, such as NIST (Schwartz et al., 2022), ISO (Joint Technical Committee ISO/IEC JTC 1, Information Technology, Subcommittee SC 38, Cloud Computing and Distributed Platforms, 2020) and IEEE (Intelligent Transportation Systems Committee of the IEEE Vehicular Technology Society and Standing Committee for Standards or the IEEE Robotics and Automation Society, 2021), work to provide broad guidance in terms of documentation over technical fields of all kinds, we expect that the answers to these questions and others for localized research communities will require active and inclusive community engagement to encourage uptake and effective documentation processes, practices and products.

On co-evolving technology and social structure Value sensitive design points us to the need and opportunity to co-evolve technology with social structure (Friedman and Hendry, 2019). That is, by developing technical tools and toolkits along with the social environments in which they will be used, we have a larger design space with which to engage and greater possibility to ensure that resulting practices will be responsive to the needs of individuals, communities, fields and society writ large. Doing this kind of co-evolution work is complex, nuanced work. Mok and Hyysalo (2018) explore such co-evolution in the context of energy transition for a historical building in Finland; Magassa and Friedman (Under review) for the Washington State Access to Justice Technology Principles. Our work improving the data statements documentation toolkit contributes a focused case study for such co-evolution — one in which we worked directly with the community of practice both to improve the technology and to explicitly identify best practices around the technology's use. Our final products reflect this co-evolution approach, resulting in both a revised documentation toolkit (Data Statements Schema Version 2) and a set of best practices and guide for writing data statements. As the data statement toolkit is integrated into community practice, these methods could be used to understand how the integration process has changed the community and how those community changes necessitate the schema be once again revised. The overall approach we have taken as well as some of our specific methods for simultaneously engaging with a community around the development of the technical artifact will be of use to others who wish to pursue such co-evolution in their own design situations.

3.8 Future Work

The approach and methods reported here make progress on the trajectory from technical concept to widespread community practice. Yet more remains to be done. We point to three promising directions for future work.

Iteration and integration: use cases and on-going technical refinement As data statements for NLP systems continue to be taken up, engaged and refined by diverse stakeholders, as a field we will be positioned to study their adoption, adaptation and effectiveness in practice. Open research questions include:

- What use cases emerge for data statements for NLP systems?
- How does the data statements schema for language data types need to be refined so as to be fully general, accommodating all kinds of observational data that may co-occur with or provide context for text or audiovisual language data?
- How does domain of use and organizational context impact the content of data statement schema elements and how those elements are used in practice (e.g., medical texts with patient, disease, and drug information vs. legal texts with case law)?
- How do diverse stakeholders read data statements and how readable are data statements, particularly for non-technical stakeholder groups?
- What evidence is there for the success of data statements for NLP (and related documentation toolkits) in mitigating bias and enabling better science?
- Where and how do data statements as a documentation toolkit come up short?

Generalizing to other data types A key strength of data statements is their precision in relation to the dataset's data type. That is, the schema elements are honed to the data type that is being documented. The strength comes at the expense of generalizability, that is how readily data statement schema elements that were initially developed for language data types as used in NLP systems could be adapted in conception and structure to other data types. Our intuition was that some elements of the schema would likely carry across to other data types. After all, documentation for any data type will need to address the reasons underlying

selection and inclusion (i.e., Curation Rationale) as well as disclosures and information on ethical review processes (i.e., Disclosures and Ethical Review). But elements specific to language data would need to be removed and new elements relevant to the data type being documented would need to be developed. Datasets with mixed data types (e.g., images with captions) present further complexities in documentation.

To explore further, we conducted a thought experiment as follows. Each of the authors chose a different (non-language) data type and considered how the schema elements developed for language data types might apply: vision data used to detect motion; sensor data used to train autonomous vehicles; and electrical signal data used in brain-machine interaction. We compared our judgments about each schema element. By consensus we identified only 4 out of the 15 schema elements that would not carry over (elements 5–8: Language Varieties, Speaker Demographic, Annotator Demographic, and Speech Situation and Text Characteristics). The remaining 11 elements all carried over to each of the three considered data types, in some cases without modification, in others with minor adaptation to the element description. A development process akin to that of data statements for NLP could build out data statements for additional data types, replacing elements 5–8 with data type specific elements and adapting the details of the others. This thought experiment suggests that the grounding of the data statements toolkit in a specific data type, far from making it inflexibly bound to that data type, produced a resource that would be a beneficial starting point for adaptation to other domains.

Engaging with a broader set of stakeholders Value sensitive design calls for a robust engagement with both direct and indirect key stakeholder groups. A stakeholder analysis for data statements yields many, diverse stakeholder groups, each of whom may interact with data statements in distinct ways. These include but are not limited to those (linguists, data scientists and others) who *create datasets*; those (computer scientists, data scientists and others) who *develop systems trained and tested on datasets* created by others; those (institutional decision-makers and IT personnel in organizations) who *select systems trained on datasets* created by others; those (doctors, human resources personnel, judges, lawyers, loan officers and others) who *use the outputs of systems trained on datasets* created by others; and those (individuals, communities, advocacy organizations and societies) who may *never touch the systems that were trained on the datasets but nonetheless are affected* by how others interpret and act upon the outcomes. All of these stakeholder groups need to be brought into the design process for data statements to ensure that the documentation contains

the necessary information to be useful and that information is presented in a readable, comprehensible and usable form and format for each of the stakeholder groups. The work presented in this chapter primarily addresses only the first stakeholder group above — those who create datasets. In the following chapters, I begin to incorporate the perspectives of some of these additional stakeholder groups, focusing on the communities whose language(s) are encoded in datasets and systems trained on those datasets.

3.9 Conclusion

Responsible approaches to machine learning will only gain purchase when the tools and technologies designed to support these outcomes are taken up and integrated into the everyday practices of technical and non-technical communities alike. In the work reported here, we explored how to support uptake of such a toolkit within in one particular technical community: data statements within the NLP community. Along the way, we also demonstrated how engagement with the technical community can be used to improve the toolkit, thus achieving two goals with one intervention. Framed in this manner, our work makes four key contributions. First, we provide a revised version of the data statements schema, together with a set of best practices for writing data statements, both presented together in a guide for writing data statements. Second, we developed a method for engaging a technical research community in uptake and adaptation of a documentation toolkit for machine learning systems, including workshop structure and interaction strategies. Third, with respect to improving the documentation toolkit itself, we provide a method and practice for further developing and improving such toolkits. Finally and most generally, we demonstrate how to move from an early-stage technical concept and innovation informed by value sensitive design to a community practice around a more robust technical artifact.

I draw from the methodology and lessons in this work to develop C3DAR. In Chapter 4, I again rely on value sensitive design methodologies to structure the next steps towards a future-looking toolkit for documenting and developing datasets with language communities. Chapter 5 presents a retrospective investigation grounded in the same coding methods we used to analyze the workshop data. The results of the retrospective technical investigation are then leveraged in combination with the datasheets analysis work that we deemed out of scope for the version 2 schema to transform the version 2 schema into the first version of C3DAR (Chapter 6). Critically, the work in the following chapters remains grounded in the practice of

iterative and cyclical development.

Chapter 4

Methodology

In this chapter, I detail the value sensitive design methods and social science methods that I draw on to develop the C3DAR toolkit. Following value sensitive design, I conducted two investigations, retrospective technical and technical, to develop C3DAR as a collaborative planning tool. I describe the methods for my first investigation, a retrospective technical investigation, in §4.1. In this investigation, I analyze ethical guidelines for research involving minoritized language communities and licenses for language technology to survey communities' published values for ethical research. In §4.2, I outline a technical investigation in which I develop two interim versions of C3DAR, C3DAR Versions 0.1 and 0.2, and the final C3DAR Version 1. Figure 4.1 shows this process for developing C3DAR as a flowchart.

4.1 Retrospective Technical Investigation

The purpose of this retrospective technical investigation is to provide a fundamental understanding of the current space of language community dataset development, including who is involved and how competing values may be prioritized. To answer research questions around how best to improve the Version 1 data statements schema from the documentation writer's perspective, we designed a workshop (see Chapter 3) to leverage the experience of our workshop participants who came from diverse research backgrounds. This helps us address the needs of NLP practitioners writing documentation for language datasets, but it still leaves open several questions. First, to limit the scope of our work while developing data statements Version 2, we refined the documentation schema to address existing datasets rather than having the documentation

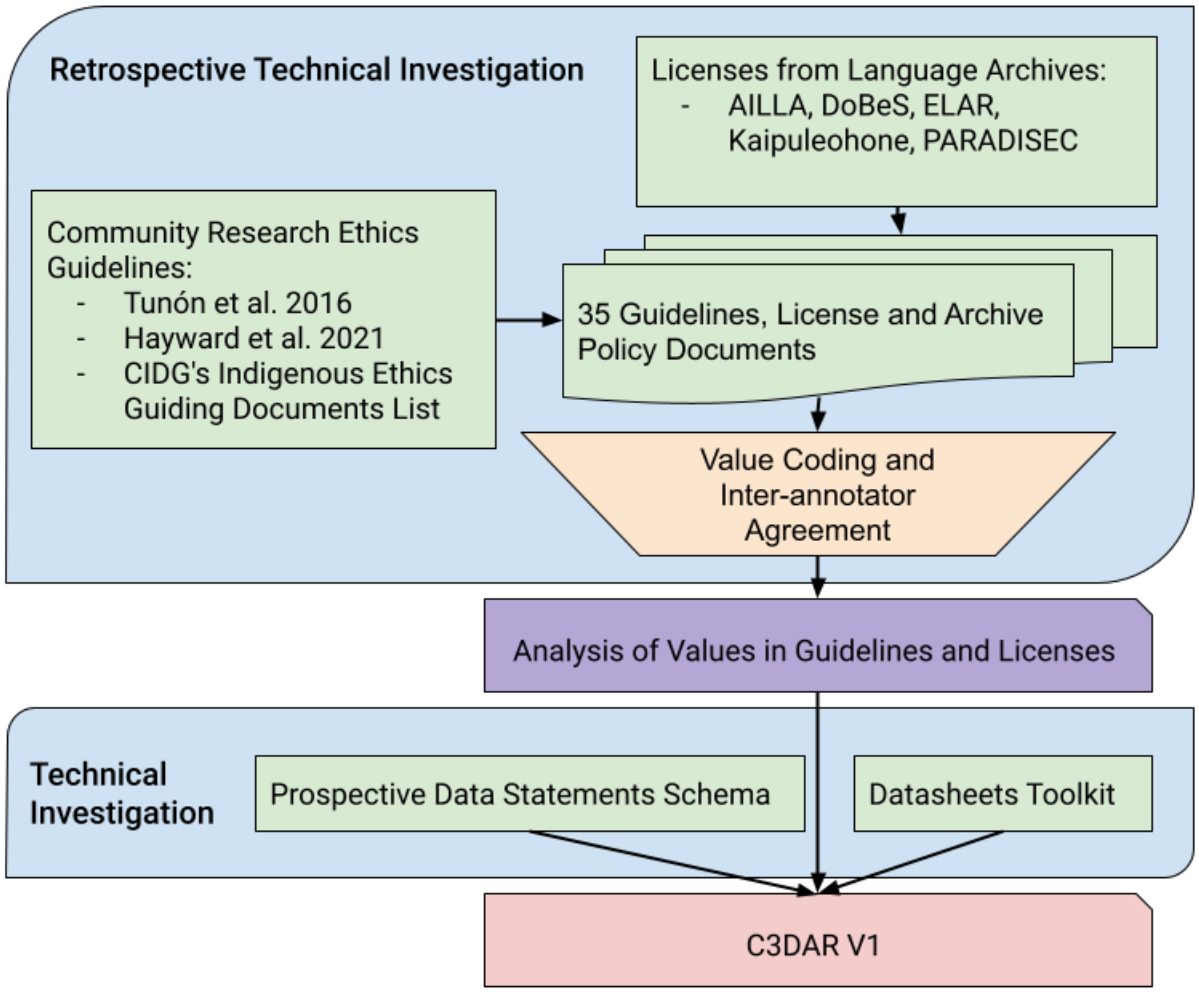

Figure 4.1: Flowchart showing the development process for C3DAR

.

serve a prospective function for future datasets. Our findings from the workshop were therefore not set up to investigate participants' perspectives on the values embedded in datasets creation and management. As a result, Version 2 currently does not include forward-looking dataset considerations such as creating datasets and ongoing dataset management. It also does not include language community perspectives on the values that are important for creating and managing datasets. This retrospective technical investigation provides the foundation for bringing those considerations into the scope of C3DAR, particularly considerations from the perspectives of language communities.

To understand how the groups involved in the dataset development cycle may differ in their prioritization of values in creating and managing language data, I start from the technology used by the groups to communicate those values, namely ethical research guidelines and technical licenses for data. Value sensitive design takes a broad stance in defining technological systems as the interdependent relationships between tools, technologies, and infrastructure that shape human activity, and from this definition value sensitive design has also come to include policy as a combination of tool, technology, and infrastructure that shapes human activity through law and regulation (Friedman and Hendry, 2019; Young et al., 2019). Licenses and ethical research guidelines are both policy tools for regulating data creation and management, but they function in different manners and have overlapping but not identical sets of intended audiences. Licenses regulate the use of data by third party users, and failure to comply with the terms of the license may be met with legal action against the user. Ethical research guidelines are less strongly tied to legal regulation, but are tied to academic expectations of adherence to ethical standards. Failure to comply with ethical guidelines may therefore result in the research being rejected or discredited by the academic community. By looking at these tools as written by both minoritized language communities and other research organizations, I gather a set of expectations for data that point to the values of both groups in creating and managing data. The result of this first retrospective technical investigation enables the next investigation (§4.2), in which I use this set of values and expectations to develop C3DAR by incorporating these perspectives on community language dataset creation and management.

To analyze these tools, I build off the work of Tunón et al. (2016) and Hayward et al. (2021). Tunón et al. (2016) sought to compare the values in ethical codes and guidelines related to indigenous knowledge and cultural practices from different authoring institutions, understand the development processes of those codes

and guidelines, and study researchers' awareness of them. They looked at 12 ethical guidelines for biodiversity developed by indigenous communities (namely the Nordic Saami Parliaments, the Assembly of First Nations, and Māori communities), the United Nations, and academic associations and institutions from Australia, Canada, Sweden, and the United States. They coded those guidelines for the 18 principles identified in the International Society of Ethnobiology (ISE) Code of Ethics (International Society of Ethnobiology, 2006 with 2008 additions) and found 5 "core" ethical principles that most of the guidelines addressed: respect, recognition of rights, responsibility as a scholar, participation, and mutual benefits. Tunón et al. also included mindfulness in their core principles as it is directly stated in the ISE Code of Ethics preamble (see §5.3 for the full list of International Society of Ethnobiology (2006 with 2008 additions) principles and their definitions). They concluded that differences between the guidelines arose from the various contexts in which they were developed and conflicts between the guidelines arose from the differing author perspectives. While Tunón et al. (2016) collected documents from diverse institutional and regional contexts, Hayward et al. (2021) identified commonalities between 20 indigenous research ethical guidelines from exclusively First Nations, Métis, and Inuit communities and organizations in Canada. Their goal was to survey indigenous communities' processes for establishing their own research guidelines and show the impact those guidelines have had on research approaches in academic and government institutions. Hayward et al. surfaced three themes across the documents: balancing individual and collective rights; upholding culturally grounded ethical principles, and self-determined research processes, methods, and knowledge translation.

Whereas Tunón et al. (2016) and Hayward et al. (2021) focused on indigenous communities, I take a broader approach and include all minoritized language communities within the scope of this project. This includes indigenous communities, signed language communities, and others whose linguistic rights have historically been oppressed and who now work to establish protections for their linguistic and cultural knowledge through various frameworks of data governance. Researchers pushing for more community-directed research for signed languages such as Harris et al. (2009), Fenlon et al. (2015), and O'Brien (2017) have drawn inspiration from literature from indigenous data sovereignty, making a connection between the historical similarities between indigenous cultures and sign language communities in establishing rights to and protections for their linguistic knowledge (see Chapter 2). I see the opportunity to extend the conversation between these two lines of research into the dataset documentation space in order to design C3DAR

to accommodate a broad range of data considerations. Just as indigenous languages and perspectives are not monolithic, so to do the varied social and political contexts of signed languages around the world provide unique considerations for data and its documentation. Additionally, Tunón et al. (2016) and Hayward et al. (2021) consider only ethical research guidelines. Including licenses in the analysis provides insight into communities' perspectives on sharing data and managing later data use which is not always covered in ethical research guidelines. With such diverse perspectives and sources of values, this investigation surfaces not only commonalities but also the tensions between the values to expand on Tunón et al.'s and Hayward et al.'s work.

4.1.1 Constructing a List of Potential Documents

To construct a list of potential documents to analyze, I start from the lists of ethical guidelines analyzed by Tunón et al. (2016) and Hayward et al. (2021) as well as the Collaboratory for Indigenous Data Governance (CIDG)'s Indigenous Ethics Guiding Documents List (David-Chavez et al., 2020; Natonabah et al., 2020). The CIDG list contains examples of research guidelines published by indigenous communities like the South African San Code of Research Ethics (South African San Institute, 2017), Alaska Federation of Natives Guidelines for Research (Alaska Native Knowledge Network, 2006), and Te Ara Tika, Guidelines for Māori Research Ethics (Hudson et al., 2010). Some of the guidelines in the CIDG list pertain to public health research and may not be relevant to the current study focused on language data, in which case I remove them from the analysis. Additionally, the lists from CIDG, Tunón et al. (2016), and Hayward et al. (2021) consist primarily of ethical research guidelines developed by indigenous communities and partnering organizations in the Americas and Oceania. To broaden these perspectives, I supplement these lists with ethics guiding documents from diverse communities such as the Sign Language Linguistics Society's Ethics Statement for Sign Language Research (Sign Language Linguistics Society, 2016) and the Sámi Council's discussion paper for Ethical Guidelines Research (Holmberg, 2021).

As discussions of data ownership and sovereignty have evolved, communities have also turned to licenses to communicate their values and expectations around the uses of their data, especially by third parties outside of the community. In addition to releasing datasets with general, multipurpose licenses such as the Creative Commons licenses, communities have also developed their own language technology licenses such

as the Kaitiakitanga License[1] developed by the Māori community (Te Hiku Media, 2022). While guidelines provide a more general perspective on the expected relationship between the community and institutional researchers, licenses state specific expectations with respect to a defined context and a known portion of community knowledge, the dataset. Guidelines also tend to focus on encouraged actions, whereas licenses also provide insight into disallowed actions such as restrictions on redistribution. For these reasons, investigating licenses as well as guidelines introduces a greater variety in the expression of community values.

To find additional licenses and ethical research guidelines, I searched through language documentation archives. These archives, operated and managed by universities and museums, support data storage for minoritized language communities around the world and are a recommended option for long-term data storage and access (Hanke and Fenlon, 2022). Data curation is costly for communities; communities can rarely afford to purchase and maintain their own data servers, and paying for server storage through a commercial data operator risks losing the data if the community is ever unable to continue payment. Long-term storage in a linguistic research repository centralizes the infrastructure costs of data storage while ensuring that the data will remain accessible to the community and affording the community control over who else can access their data. Communities who choose to store their data in these archives have therefore implemented data curation and management processes; further documentation of the final processes, either as guidelines or licenses, may be stored in the archive in the form of metadata.

For each archive, I searched the collections and compiled a list of projects that contain such licensing or guideline documents. Rather than searching the archives exhaustively, I randomly searched 50 collections from each archive, expecting that many will have insufficient prose metadata to be used in the later analysis. Additionally, these archives also have their own policy documents which I included in the analysis. I searched the following archives:

- Archive of Indigenous Languages of Latin America (AILLA) - University of Texas at Austin, USA[2]
- DoBeS archive - Max Planck Institute in Nijmegen, the Netherlands[3]
- Endangered Language Archive (ELAR) - Berlin-Brandenburg Academy of Sciences and Humanities

[1]Permission to include the Kaitiakitanga License in this research received by email January 8th, 2023.
[2]`https://ailla.utexas.org/`
[3]`https://dobes.mpi.nl/`

in Berlin, Germany[4]

- Kaipuleohone - University of Hawai‘i at Mānoa, USA[5]
- Pacific and Regional Archive for Digital Sources (PARADISEC) - University of Sydney, Australia[6]

The full set of collected documents is detailed in §5.1.

4.1.2 Selecting a Document Set for Analysis

I aimed for the analysis to include five English-language[7] documents each from Africa, North America, South America, Asia, Europe, and Oceania to ensure diverse perspectives are included in the analysis. Within each region, I selected documents based on three properties: 1) the author of the document (whether the document was authored by a minoritized language community or by an outside research or government organization); 2) the type of document (ethical research guideline or license); and 3) the modality of the community's language (spoken or signed). The document sets for each region were selected to have four documents each authored by minoritized language communities as well as one document authored by research or government organization. Each set was also selected in order to include at least one ethical guideline document and at least one licensing document. Finally, each region includes the perspective of at least one signed language. Table 4.1 shows a schematization of three possible document sets for a region, where each possible set adheres to these minimum document property requirements. Table 4.2 also shows three hypothetical document sets which would be disallowed due to failure to meet one of the property requirements. The failed property requirement is shown in red. Set 4 has two documents written by a research or government institution, Set 5 has only guidelines, and Set 6 has only documents for spoken languages. When I collected more than the minimum number of documents, I selected documents at random such that they satisfy the property requirements for a given region. This results in 30 total documents plus the 5 archive policy documents (see §5.2 for the full list of documents).

[4] https://www.elararchive.org/

[5] http://ling.hawaii.edu/kaipuleohone-language-archive/

[6] https://www.paradisec.org.au/

[7] The primary reason for choosing documents published in English was for my own ease of access. Secondarily, English is commonly used in academic settings, so documents addressed to international researchers are likely to be published in English. That being said, English is not the working language in many contexts. This sampling strategy will therefore exclude documents published in the community's language as well as documents published in more widely spoken languages and national languages that are not English.

Set	Properties	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Possible Set 1	Author	Community	Community	Community	Community	R/Gov
	Type	Guide	License	Guide	Guide	Guide
	Modality	Signed	Spoken	Spoken	Spoken	Spoken
Possible Set 2	Author	Community	Community	Community	Community	R/Gov
	Type	License	License	Guide	License	License
	Modality	Signed	Spoken	Spoken	Spoken	Spoken
Possible Set 3	Author	Community	Community	Community	Community	R/Gov
	Type	Guide	License	Guide	Guide	Guide
	Modality	Spoken	Signed	Signed	Signed	Signed

Table 4.1: Example possible region document sets. Each set has 4 documents written by communities and one by a research or government institution; at least 1 guide and at least 1 license; and at least 1 spoken and at least 1 signed language represented.

Set	Properties	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Disallowed Set 4	Author	Community	Community	Community	R/Gov	R/Gov
	Type	Guide	Guide	License	Guide	License
	Modality	Spoken	Signed	Spoken	Spoken	Signed
Disallowed Set 5	Author	Community	Community	Community	Community	R/Gov
	Type	Guide	Guide	Guide	Guide	Guide
	Modality	Spoken	Signed	Spoken	Spoken	Signed
Disallowed Set 6	Author	Community	Community	Community	Community	R/Gov
	Type	Guide	Guide	Guide	Guide	License
	Modality	Spoken	Spoken	Spoken	Spoken	Spoken

Table 4.2: Disallowed document sets with the disqualifications in red. Set 4 has two documents written by a research or government institution. Set 5 has only guides. Set 6 has only documents for spoken languages.

4.1.3 Value Identification and Coding

To learn about the distribution of values in community language documentation, I coded each sentence in the documents for the values they reflect, using the ISE Code of Ethics (International Society of Ethnobiology, 2006 with 2008 additions) also used by Tunón et al. (2016) (see §5.3). Tunón et al. analyzed their own set of ethical guidelines from international, academic, and community perspectives using the principles in the ISE Code of Ethics as the point of comparison across the guidelines. The ISE Code of Ethics was developed for ethnobiologists in collaboration with indigenous and non-indigenous groups from around the world. As an international framework, the ISE Code of Ethics could be assumed to have taken into account broad understandings of values and principles across many different indigenous communities. Because Tunón et al. (2016) were investigating other guidelines for ethnobiological and general research with indigenous communities, the ISE Code of Ethics principles served their purposes without any changes. As my investigation method includes coding guidelines (as opposed to comparing them), analyzing licenses, bringing in perspectives from signed language communities, and addressing both academics and community members (rather than just the academic audience of the ISE Code of Ethics), I first needed to adapt the ISE Code of Ethics into a coding manual with considerations for licenses, signed language communities, and language communities audiences. I also was not focused on ethnobiological research (though some of my guidelines were centered on ethnobiological research), so the coding manual needed to have a general research perspective to account for the perspectives in the documents I collected.

To adapt the ISE Code of Ethics for use as a coding manual, I first used the principles as presented in the ISE Code of Ethics and an additional *None* label to code half of the documents selected at random for the values reflected in each sentence. I then changed the coding manual text and labels to better represent the values I saw in the first 15 documents. I also added labels to include values that the original text did not cover. I then coded all of the documents, including the first 15 used to adapt the coding manual, using the revised coding manual.

An advantage to starting from the ISE Code of Ethics to develop my coding manual was the strong foundation in indigenous principles and collaboration between indigenous communities and research partners. I found support for all of the original labels in the 15 documents I first coded. On the one hand the mismatches between the ISE Code of Ethics and my investigation were a disadvantage because the ISE Code of

Ethics principles were developed to support one another, as principles are often connected with each other or actions toward one principle may result in support for another principle. For a coding manual, however, labels need to be contrastive in order for the labels to be systematically applied. Disentangling the ISE Code of Ethics principles in the coding manual proved to be a significant challenge and impacted the results found in Chapter 5. On the other hand because of the mismatches, my process required coding some of the documents twice. I found that the process of coding the documents once to develop the coding manual and again to code the documents with the finalized coding manual gave me more practice with the coding exercise and allowed me to be more systematic in my annotations than if I had only analyzed each document once.

After I finished coding all of the documents using the coding manual, a second annotator, Dr. Emily M. Bender, separately coded 10% of the documents selected at random. I calculated Cohen's Kappa (Cohen, 1960) to show the extent to which I and Dr. Bender agreed on the values represented in each sentence of the documents (§5.3.3). This investigation resulted in a list of values in community language documentation from around the world (§5.3.1), relative distributions of the values by geographic region and document types (§5.3.2), and examples of the phrases used to convey those values (§5.4). From these results, I surfaced interactions between the values in the processes of dataset creation and documentation in the following technical investigation.

4.2 Technical Investigation

To design a toolkit that is responsive to both language communities' and technical communities' needs while creating language datasets, I started from the Version 2 data statements schema. While other documentation schemas might also be suitable starting points in that they serve a variety of technical audiences, data statements are the only currently available documentation format that are explicitly developed with language as the intended data type being documented. The schema therefore contains elements that are already designed to address the unique considerations for documenting language data. In Chapter 3, my co-authors and I refined the data statements schema to better serve technical communities' needs for documenting datasets. I further refined the schema to 1) shift the perspective of the data statements Version 2 from documenting existing datasets to planning datasets, 2) incorporate the needs of language communities, and 3) support collaboration between technical communities and language communities.

I developed project in three stages. The first stage of this process required changing the perspective of the data statements Version 2 from documenting existing datasets to planning datasets, meaning I changed references to the dataset being documented from present tense to future tense and references to the dataset creation process from past tense to future tense. These changes produced C3DAR Version 0.1. To produce C3DAR Version 0.2, I incorporated maintenance and distribution considerations from datasheets for datasets (Gebru et al., 2021) as new schema elements, discussed further in §4.2.1. Finally, I made changes to align C3DAR with the recommendations from the guidelines and licenses analyzed in the retrospective technical investigation throughout the C3DAR schema, resulting in C3DAR Version 1 (§4.2.2).

4.2.1 Incorporating Elements from Datasheets

The responses from the workshop participants (§3.5.1) suggested that the data statements schema may be helpful determining what information for documentation will need to be collected at the design phase of a dataset. We incorporated this insight as General Best Practices 3 and 4, however, in further developing the Version 2 data statements schema, we made the explicit decision that writing and publishing finalized data statements is a retrospective exercise with respect to the dataset itself, rather than an exercise supporting the design of the dataset. One may begin drafting the data statement before the dataset is complete, but published data statements are expected to describe up-to-date information about existing datasets. This affected the results of our technical investigation comparing data statements to datasheets in §3.5.2. We chose not to incorporate several of the questions in datasheets into the data statements schema because they ask for prospective information, especially in the Distribution and Maintenance sections of the datasheet framework.

I added the majority of the content from the Distribution and Maintenance sections from datasheets into the Version 2 data statements schema. Some of the content, however, is not applicable to language data and community dataset creation as datasheets were developed for general purpose machine learning dataset documentation in an industry context. In such cases, I edited the language to be more relevant to the language community context. Furthermore, the style needed to be edited to match the other data statement elements. The sections from datasheets contain questions grouped by the topic, whereas the data statements elements each consist of an initial *Why* section explaining why the element is included, a *What* section

stating the content that should be provided in the final element draft, and a *Best Practices* section providing tips for drafting the element. I translated the datasheets content into a data statements *What* section, and then drafted my own *Why* for each element. I drafted the *Best Practices* sections using the results of the retrospective technical investigation.

4.2.2 Incorporating the Results of the Retrospective Technical Investigation

To start, I replaced linguistic terminology that positioned spoken languages as the norm, *speaker* and *speech* when used to refer to general language users or language use, with phrases that increased the visibility of signed languages throughout C3DAR. Depending on the context, I alternated between using the terms *language user* or *linguistic*, which does not privilege the spoken modality as the general modality, and listing out both *speaker and signer* or both *speech and sign*. Then for each value from the retrospective technical investigation, I aligned content in the schema elements or the best practices with the recommendations from the guidelines and licenses. For example, the first schema element currently requests information about authorship in a way that aligns with *Acknowledgement and Due Credit*, but may be broadened to include other kinds of contributions besides co-authorship, such as the outside participant in our previously identified interviewing technique for writing more inclusive documentation. I also used the examples to add community-oriented motivations to the *Why* sections. I used the text from the ethical guidelines and licenses to suggest new content and best practices to elements that may be improved with value-oriented statements. This included developing a new set of general best practices for collaboration; I renamed the general best practices originally developed in Chapter 3 to "general best practices for documentation" to differentiate between the two sets. In such cases where tensions between the values were foreseeable in the best practices, I noted the possible tension but left suggestions for resolving those tensions to explore in future investigations. C3DAR Version 1, presented in §6.3, contains 12 general best practices for collaboration, 15 general best practices for documentation, and 17 schema elements.

4.3 Summary

In this chapter I have laid out my methodology for designing a dataset planning and documentation toolkit using value sensitive design's iterative and integrative tripartite methodology. §4.1 introduces my retro-

spective technical investigation. I conduct this investigation in Chapter 5 by analyzing community research guidelines and licenses to aggregate communities' stated research values. In §4.2, I presented my methods for iteratively integrating data statements Version 2, datasheets for datasets, and the lessons of the retrospective technical investigation, ultimately resulting in C3DAR Version 1. The outcomes of these methods are shown in Chapter 6.

This methodology puts forward a plan for developing the first version of C3DAR based on a broad survey of communities' values. To learn more about the values and experiences held by specific stakeholders from unique communities, in Chapter 8 I propose a future empirical investigation for engaging with community dataset creation groups, representing signed and spoken language communities. In Chapter 9, I suggest a potential second technical investigation as future work to integrate the perspectives from the empirical investigation into C3DAR.

Chapter 5

Retrospective Technical Investigation: Lessons from Community Ethical Guidelines and Licenses

To begin understanding the needs of communities, I conducted a retrospective technical investigation of ethical guidelines and licenses from communities around the world. I started from the assumption that recommended and discouraged actions within ethical guidelines and licenses are grounded in values that are important to the authoring community. My goal was to surface a wide variety of values by grouping together similar recommendations from across guidelines and licenses from many different communities.

In §5.1, I provide an overview of the document collection process, resulting in 40 documents from prior work investigating ethical guidelines and 30 documents from archives with community language documentation projects. From this list, I randomly selected 30 documents to analyze; I describe the selection criteria and the documents in §5.2. I then coded each sentence in the selected documents for the values they reflect, using the list of values used by Tunón et al. (2016), that is the International Society of Ethnobiology Code of Ethics (International Society of Ethnobiology, 2006 with 2008 additions). The next section (§5.3) details the revisions to the manual after coding half of the documents, general trends in my annotations using the revised manual, and the inter-annotator agreement between myself and a second annotator, Dr. Emily M. Bender, on 10% of the documents, also using the revised coding manual. Finally in §5.4, I present important

topics surfaced from the annotation, with a selection of examples from the documents.

5.1 Document Collection

As detailed in §4.1.1, ethical guidelines and licenses serve different purposes and audiences in communicating community expectations and values. Ethical guidelines address researchers who may be engaging with the community in a wide variety of methods and focus on the relationship between the researcher and the community. Licenses, on the other hand, address a general audience in the more limited context of the allowed and disallowed uses of a particular dataset. Prior research has focused on ethical guidelines to investigate stated community values (Tunón et al., 2016; Hayward et al., 2021); as more communities have turned to licenses as a method for protecting data, adding licenses to the potential documents under investigation presents the opportunity to understand the community's values in a different context.

My goal was to analyze diverse, English-language guidelines and licenses across geographic regions and along three properties: 1) the author of the document (whether the document was authored by a minoritized language community or by a research or government organization); 2) the type of document (ethical research guideline or license); and 3) the modality of the community's language (spoken or signed). I aimed to select 5 documents from Africa, North America, South America, Asia, Europe, and Oceania for a total of 30 documents. Within each of these groups of 5 documents, I outlined minimum requirements across the three properties above: four documents in the set must be authored by minoritized language communities, with the fifth document authored by research or government organization; the set must include at least one ethical guideline document and at least one licensing document; and at least one document will focus on a signed language community. For examples of hypothetical document sets meeting and failing to meet these requirements, see Table 4.1 and Table 4.2.

I started from a list of ethical guidelines collected by Tunón et al. (2016) and Hayward et al. (2021). I searched for additional ethical guidelines for indigenous and signed language communities using Google Search and by examining referenced guidelines and organizations from the documents I had collected so far. Once that search was completed, I collected licenses of community language documentation datasets from five archives, listed in §5.1.2, until I had sufficient documents to meet the geographical and document property requirements. These archives publish policy documents pertaining to researcher conduct and data

rights, so I included those documents in the analysis as well. Altogether I searched through more than 400 documents to collect a group of 71 documents that I could potentially analyze. The next sections describe how I found these 400 documents from prior work and in archives.

5.1.1 Collecting Documents from Prior Work

In §4.1, I summarized the goals and methods of Hayward et al. (2021) and Tunón et al. (2016). Hayward et al. (2021) analyzes 20 ethical guidelines, all from indigenous communities in North America, specifically Canada. Tunón et al. (2016) analyzes 12 guidelines. Of these three are international documents, three are from Europe, four are from North America, and two are from Oceania. Tunón et al. (2016) also note the institutions the authors of the guidelines are affiliated with: four community authored guidelines and eight guidelines from research or government organizations. All of the guidelines from both sources focus on spoken language communities.

In reviewing the work of the authors of the CARE principles (Carroll et al., 2020) and in particular Dr. Stephanie Russo Carroll's work in indigenous data governance, I came across the Collaboratory for Indigenous Data Governance and their project with the Indigenous Land and Data Stewards Lab[1] on developing a national standard for indigenous ethics research (David-Chavez et al., 2020; Natonabah et al., 2020). As part of this effort, David-Chavez et al. (2020) collected 31 ethics documents developed by indigenous Nations, communities, and working groups.[2] The majority of the documents are from North America, but the list also includes documents from South America, Oceania, Africa, and international teams, all for spoken language communities. I considered the documents from this list as well in collecting documents for analysis.

From the 63 documents I collected in this way from prior work, 20 were removed from consideration for this particular investigation. In one case, the link to the document was no longer active and searches for the document failed, and in another the original document, titled "Nordic Saami Convention," was in Swedish with no official English translation available. I replaced this document with Holmberg (2021), an English-language document discussing the Sámi Council's work towards ethical guidelines for research concerning Sámi people which cites Tunón et al. (2016). Three documents from Tunón et al. (2016) and four from David-Chavez et al. (2020) also included documents developed by international teams, like the UNDRIP

[1] `https://www.indigenouslandstewards.org/`

[2] The website states 32 documents, but lists 31.

and the CARE principles, and so were removed as they would not fit within a specific region category. Five of the documents reviewed by Hayward et al. (2021) are tailored to guidance for health research, so those were removed as this work is concerned with language research. Two documents from Hayward et al. (2021) and four documents from David-Chavez et al. (2020) were removed because they were not documents for research guidelines, but rather templates or guides for establishing research policies. Finally three documents were duplicated between two or more of the sources, resulting in a final set of 41 guidelines from indigenous perspectives. For the final set of included documents, I noted the properties of the documents according to author type, document type (guideline), and language modality.

Additionally, in anticipation of not being able to find enough documents from signed language communities, I collected several guidelines from academic and international signed language perspectives to be included only in the case that I couldn't satisfy the signed language perspective requirement for each geographic region. In searching online for "signed language research ethics," the Sign Language Linguistics Society (SLLS) *Ethics Statement for Sign Language Research* (Sign Language Linguistics Society, 2016) was one of the first returned sites. The majority of the search returns, however, focused on interpretation ethics. For that reason I focused on the work cited by and citing the SLLS Ethics Statement. The SLLS Ethics Statement refers readers to Harris et al. (2009), which presents the six principles of the sign language communities' terms of reference (SLCTR). The SLLS Ethics Statement also suggests the World Federation of the Deaf (WFD) and the Finnish Association of the Deaf *Manual for Sign Language Work within Development Cooperation*, specifically Chapter 6 "Best Practices and Challenges in Sign Language Work" (World Federation of the Deaf and Finnish Association of the Deaf, 2015), for further information on sign language ethics. This publication is itself cited by the WFD *Best Practices and Ethics for Development Co-operation Projects* (World Federation of the Deaf Expert Group on Developing Countries, 2016). I collected these four documents as reserve documents to ensure I would be able to include at least five signed language documents, with the assumption that I would find at least one signed language license in the archives.

Finally, I intentionally included the Te Hiku Media Kaitiakitanga License[3] developed by the Māori community (Te Hiku Media, 2022). Dr. Bender pointed me to Te Hiku Media early in my review of the indigenous data sovereignty literature. The Kaitiakitanga License was the inspiration for looking for other

[3]Permission to include the Kaitiakitanga License in this research received by email January 8th, 2023.

community data licenses and investigating how other communities have established protections for their data. Initial internet searches for licenses proved unsuccessful, however, so I turned to language data archives as the canonical platform in linguistic research for disseminating language datasets and documentation projects.

5.1.2 Collecting Documents from Archives

Archives for language documentation contain language data along with various amounts of linguistic, anthropological, and/or sociological documentation and metadata, depending on the depositor. Some are historic collections from as early as the 1950s that have been received by the archive and then digitized to make them broadly available. These tend to have less documentation and fewer descriptions about the community; more information is typically available about the researcher who originally created the recordings. The newer collections are more likely to be the result of collaborations between researchers and communities, with communities aligning research agendas with their own needs. In addition to providing long-term storage and centralizing the costs of the hardware and technical support, the archives also provide communities with control of their data through systems of access protocols, which vary by archive.

I chose five archives to search for potential documents.[4] The Archive of Indigenous Languages of Latin America (AILLA)[5] is hosted by the University of Texas at Austin, USA. AILLA contains 270 collections documenting communities from North and South America. The Language Archive[6], hosted by the Max Planck Institute in Nijmegen, the Netherlands, contains multiple repositories specializing in different kinds of language data. For this work, I focused on the DoBeS (Dokumentation bedrohter Sprachen) Archive[7] and the Sign Language archive. The DoBeS Archive holds 71 collections from around the world, and the Sign Language archive holds 12 collections from Europe, Africa and North America. The Endangered Language Archive (ELAR)[8] is hosted by the Berlin-Brandenburg Academy of Sciences and Humanities in Berlin, Germany. ELAR is the largest of the archives and has the greatest regional variety with 738 collections from around the world. Kaipuleohone[9], meaning "gourd of sweet words" in Hawaiian, is the language archive of the University of Hawai'i at Manoa, USA. Kaipuleohone's 78 collections primarily contain community data

[4]The number of collections for each archive was counted in February 2023.
[5]`https://ailla.utexas.org/`
[6]`https://archive.mpi.nl/tla/`
[7]`https://dobes.mpi.nl/`
[8]`https://www.elararchive.org/`
[9]`http://ling.hawaii.edu/kaipuleohone-language-archive/`

from Oceania and Asia, though also some from the Americas. The fifth archive is the Pacific and Regional Archive for Digital Sources (PARADISEC)[10]. It is the second largest archive with 544 collections, largely from Oceania and Asia.

An exhaustive search of the archives would be unnecessarily time-consuming; to limit the scope of my search, I planned to search 50 random entries from each archive. To do this random search, for each archive I assembled a list of all of the collections in that archive. The organization of the list was determined by the organization by the archive, usually alphabetical, but with some archives organizing by language and others by collection title. I then used a randomization function in Google Sheets to randomize the entries in each list and investigated the top 50 entries of each list for each archive.

To inform my later document selection, I noted properties about each of the collections I investigated. The first category was whether the collection had sufficient relevant metadata to be useful in my analysis. Relevant metadata included comments about intended use of the data and details about the creation of the data collection. Some collections discussed only the research goals of the data collector or had no prose descriptions at all. From the collections with relevant metadata, I noted the three properties I outlined at the beginning of §5.1: the author type, the document type (license), and the language modality.

After investigating 50 collections from each archive, I assessed how many entries I had noted with sufficient metadata per region and modality. At this point, I had found 28 collections with sufficient metadata, but I was still missing signed languages from Africa, Asia, North America, South America and a guideline document for Asia. I continued to search for documents in ELAR as collections I had encountered in ELAR had the most relevant prose documentation of the archives. With the intent to look for signed languages, I used their search function to find all collections with "sign" in the language name. This turned up a signed language collection from Africa and another from Asia, but not North America and South America (with enough relevant documentation), bringing my total count to 30 collections. Satisfied with this search, I resolved to use my reserved international signed language documents to supplement the North and South America document sets and to satisfy the guideline requirement for the Asia document set.

[10] `https://www.paradisec.org.au/` hosted by the University of Sydney, Australia

5.1.3 Description of Collected Documents

For both the prior work and the archives, I coded the documents for the region as well as the minimum requirement properties: 1) the author of the document (whether the document was authored by a minoritized language community or by a research or government organization); 2) the type of document (ethical research guideline or license); and 3) the modality of the community's language (spoken or signed). While the document type and the modality of the language were straightforward to determine, the author type required more effort. I relied on stated affiliations (e.g., with a specific community) or comments on the document author's relationship with the community. These included both direct statements (i.e., "I am a member of the community") and more indirect comments (e.g., using the third person when referring to the data or community and referring to the community members as "collaborators" or "consultants"). When in doubt, I labeled the author as a non-community researcher because this is more likely to be the case when the affiliation of the author is not stated.[11]

Of the 41 guidelines collected in §5.1.1, two were from Africa, three from Europe, 30 from North America, five from Oceania, and one from South America, all representing spoken languages. Of those I determined 23 to be from communities and 18 from research or government organizations. I then collected an additional four signed language guidelines from international and research communities and one community license from Oceania.

Of the 30 licenses that could be analyzed from the archives, 16 were from ELAR, 11 from The Language Archive (DoBeS and the Sign Language archive), two from AILLA, and one from Kaipuleohone. The region distribution included eight from South America, seven from Asia, five from Europe, four from Oceania, three from Africa, and three from North America. Six of the licenses were for signed languages, and of the 30, 21 licenses were collaboratively developed with communities. AILLA and Paradisec had a fair amount of historic data collections with little to no metadata and more recent collections that only stated the access protocols with no description. The collections in Kaipuleohone were largely from linguistic dissertations, so the descriptions focused on the research questions and methods of the dissertation.

[11] This is based on the assumption that research teams with community partners are following the principle of Acknowledgement and Due Credit and clearly stating when publication authors have community affiliations.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
AF1. Documentation and description of a sign language in Côte d'Ivoire (Tano, 2013)	Archive	Research	License	Signed
AF2. Multimedia Documentation of Babanki Ritual Speech (Akumbu, 2014)	Archive	Community	License	Spoken
AF3. The Khoikhoi Peoples' Rooibos Biocultural Community Protocol (National Khoisan Council & Cederberg Belt Indigenous Farmers Representatives, 2019)	Prior	Community	Guideline	Spoken
AF4. Sakun (Sukur) Language Documentation (Thomas, 2014)	Archive	Community	License	Spoken
AF5. San Code Of Research Ethics (South African San Institute, 2017)	Prior	Community	Guideline	Spoken

Table 5.1: Documents selected from Africa.

5.2 Document Selection

With all of the potential documents collected together, I grouped them by region and again randomly shuffled the lists to select the top five documents for analysis such that the three feature requirements were met. If the top five documents did not meet the requirements, I disqualified the lowest document on the list and selected the next on the list that satisfied the requirement. For example, if documents 1, 3, 5, and 6 were all from research organizations, documents 3, 5, and 6 were disqualified (as only one research document could be included in each region set), and documents 7 and 8 (from communities) were included instead. The selected documents for Africa are shown in Table 5.1, for Asia in Table 5.2, for Europe in Table 5.3, for North America in Table 5.4, for Oceania in Table 5.5, and for South America in Table 5.6. The exceptions to the five-document minimum were for Asia, for which I had no guidelines, and North and South America, for which I had no signed language documents. To replace these, I randomly selected three of the four reserved signed language guidelines, shown in Table 5.7. I also included the policy documents for each of the archives (Table 5.8). These covered the terms and conditions for uses of the collections in the archive as well as codes of conduct for collection depositors.

To refer to the selected documents (not including the archives) throughout the remainder of this chapter, I assigned each document a unique ID. This ID consists of the first two letters of the geographic region that the document is selected from and a number. For example, **AF1** refers to Tano (2013) which is included in the documents selected from Africa.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
AS1. DoBeS Archive for the Waima'a Language (Belo et al., 2002-2006)	Archive	Community	License	Spoken
AS2. Documentation and Description of the Hrusso Aka Language of Arunachal Pradesh (D'Souza, 2015)	Archive	Community	License	Spoken
AS3. Preliminary Documentation of Macau Sign Language (Sze, 2014)	Archive	Research	License	Signed
AS4. DoBeS project for the documentation of the Totoli language (Leto et al., 2005-2010)	Archive	Community	License	Spoken

Table 5.2: Documents selected from Asia.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
EU1. DGS (German Sign Language) Corpus License Conditions (Konrad et al., 2020)	Archive	Community	License	Signed
EU2. DoBeS Archive for Kola Saami (Rießler et al., 2005-2017)	Archive	Community	License	Spoken
EU3. Corpus NGT (Sign Language of the Netherlands) Release Notes (Crasborn et al., 2006-2017)	Archive	Research	License	Signed
EU4. Working towards ethical guidelines for research involving the Sámi (Holmberg, 2021)	Prior	Community	Guideline	Spoken
EU5. The Preservation of the Vlaški and Žejanski Language Project (Vrzić et al., 2023)	Archive	Community	License	Spoken

Table 5.3: Documents selected from Europe.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
NA1. Catalogue of Nahuatl recordings and transcriptions from the municipality of Cuetzalan del Progreso, Puebla, Mexico (Amith, 2020)	Archive	Community	License	Spoken
NA2. Protocol and Best Practice for the Research on and Public Distribution of Information from Projects involving Indigenous Peoples (Coeur d'Alene Tribe of Idaho and University of Idaho, 2015)	Prior	Community	Guideline	Spoken
NA3. USAI Research Framework (Ontario Federation of Indigenous Friendship Centres, 2016)	Prior	Research	Guideline	Spoken
NA4. Standard of Conduct for Research in Northern Barkley and Clayoquot Sound Communities (Clayoquot Alliance for Research, Education and Training, 2005)	Prior	Community	Guideline	Spoken

Table 5.4: Documents selected from North America.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
OC1. Auslan Corpus (Johnston, 2008; Johnston et al., 2023)	Archive	Community	License	Signed
OC2. Desert Knowledge CRC Protocol for Aboriginal Knowledge and Intellectual Property (Desert Knowledge Cooperative Research Centre, 2007)	Prior	Research	Guideline	Spoken
OC3. Kani'aina Conditions of Access (Kimura (depositor), 2018; Ka Haka 'Ula O Ke'elikolani College of Hawaiian Language, 2019)	Archive	Community	License	Spoken
OC4. Kaitiakitanga License (Te Hiku Media, 2022)	Prior	Community	License	Spoken
OC5. Te Ara Tika Guidelines for Maori Research Ethics (Hudson et al., 2010)	Prior	Community	Guideline	Spoken

Table 5.5: Documents selected from Oceania.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
SA1. DOBES Archive for the Aché Language (Rößler et al., 2008-present)	Archive	Community	License	Spoken
SA2. Yman har ma'e pandu ha: Myths and accompanying co-speech gestures in Ka'apor (Godoy and Sanches de Abreu, 2022)	Archive	Research	License	Spoken
SA3. Community Biocultural Protocols (Argumedo et al., 2011)	Prior	Community	Guideline	Spoken
SA4. Tsafiki Documentation Project (Aguavil Calazacón et al., 1983-2012) and Language and Culture Archive of Ecuador (Language and Culture Archive of Ecuador, 2010a,b)	Archive	Community	License	Spoken

Table 5.6: Documents selected from South America.

Document ID, Title, and Citation	Source	Author Type	Doc Type	Modality
SL1. Best Practices and Ethics for Development Co-operation Projects (World Federation of the Deaf Expert Group on Developing Countries, 2016)	Prior	Community	Guideline	Signed
SL2. Sign Language Communities' Terms of Reference (SLCTR) (Harris et al., 2009)	Prior	Community	Guideline	Signed
SL3. Ethics Statement for Sign Language Research (Sign Language Linguistics Society, 2016)	Prior	Community	Guideline	Signed

Table 5.7: Additional international and academic signed language documents.

Archive	Document(s) and Citation
Archive of Indigenous Languages of Latin America (AILLA)	Access (Archive of the Indigenous Languages of Latin America, 2017a), AILLA Conditions of Use (Archive of the Indigenous Languages of Latin America, 2017b), AILLA License (Archive of the Indigenous Languages of Latin America, 2017c)
Endangered Language Archive (ELAR)	How to use the archive (Endangered Languages Archive, 2023)
The Language Archive (DoBeS and Sign Language Archive)	Code of Conduct (Wittenburg, 2005a), Data Access and Protection Rules (Wittenburg, 2005b), Depositor-Archivist Agreement (Wittenburg, 2005c)
Kaipuleohone	Conditions of Access (Kaipuleohone, 2008), Deposit Form (Kaipuleohone, 2015)
Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	Deposit Form and Conditions of Access (Pacific And Regional Archive for Digital Sources in Endangered Cultures, 2014)

Table 5.8: Policy documents from the archives.

5.3 Document Coding

To surface a wide variety of values, I coded the selected documents for recommendations and best practices in support of community values. For my coding manual, I started from the International Society of Ethnobiology (ISE) Code of Ethics (International Society of Ethnobiology, 2006 with 2008 additions). The ISE Code of Ethics was created to support research committed to collaborating with indigenous communities. The executive summary of the Code of Ethics "affirms the commitment of the ISE to work collaboratively, in ways that: support community-driven development of Indigenous peoples' cultures and languages; acknowledge Indigenous cultural and intellectual property rights; protect the inextricable linkages between cultural, linguistic and biological diversity; and contribute to positive, beneficial and harmonious relationships in the field of ethnobiology" (International Society of Ethnobiology, 2006 with 2008 additions). As an international framework, the ISE Code of Ethics could be assumed to have taken into account broad understandings of values and principles across many different indigenous communities. Tunón et al. (2016) used the principles (reproduced from the ISE website) from the ISE Code of Ethics to compare principles across their own set of ethical guidelines from international, academic, and community perspectives. The text of the ISE Code of Ethics is as follows:

1. Mindfulness (Preamble): “The concept of ‘mindfulness’ is an important value embedded in this Code, which invokes an obligation to be fully aware of one’s knowing and unknowing, doing and undoing, action and inaction.”

2. Principle of Prior Rights and Responsibilities: “This principle recognises that Indigenous peoples, traditional societies, and local communities have prior, proprietary rights over, interests in and cultural responsibilities for all air, land, and waterways, and the natural resources within them that these peoples have traditionally inhabited or used, together with all knowledge, intellectual property and traditional resource rights associated with such resources and their use.”

3. Principle of Self-Determination: “This principle recognises that Indigenous peoples, traditional societies and local communities have a right to self-determination (or local determination for traditional and local communities) and that researchers and associated organisations will acknowledge and respect such rights in their dealings with these peoples and their communities.”

4. Principle of Inalienability: “This principle recognises the inalienable rights of Indigenous peoples, traditional societies and local communities in relation to their traditional territories and the natural resources (including biological and genetic resources) within them and associated traditional knowledge. These rights are collective by nature but can include individual rights. It shall be for Indigenous peoples, traditional societies and local communities to determine for themselves the nature, scope and alienability of their respective resource rights regimes.”

5. Principle of Traditional Guardianship: “This principle recognises the holistic interconnectedness of humanity with the ecosystems of our Sacred Earth and the obligation and responsibility of Indigenous peoples, traditional societies and local communities to preserve and maintain their role as traditional guardians of these ecosystems through the maintenance of their cultures, identities, languages, mythologies, spiritual beliefs and customary laws and practices, according to the right of self-determination.”

6. Principle of Active Participation: “This principle recognises the crucial importance of Indigenous peoples, traditional societies and local communities to actively participate in all phases of research and related activities from inception to completion, as well as in application of research results. Active

participation includes collaboration on research design to address local needs and priorities, and prior review of results before publication or dissemination to ensure accuracy of information and adherence to the standards represented by this Code of Ethics."

7. Principle of Full Disclosure: "This principle recognises that Indigenous peoples, traditional societies and local communities are entitled to be fully informed about the nature, scope and ultimate purpose of the proposed research (including objective, methodology, data collection, and the dissemination and application of results). This information is to be given in forms that are understood and useful at a local level and in a manner that takes into consideration the body of knowledge, cultural preferences and modes of transmission of these peoples and communities."

8. Principle of Educated Prior Informed Consent: "Educated prior informed consent must be established before any research is undertaken, at individual and collective levels, as determined by community governance structures. Prior informed consent is recognised as an ongoing process that is based on relationship and maintained throughout all phases of research. This principle recognises that prior informed consent requires an educative process that employs bilingual and intercultural education methods and tools, as appropriate, to ensure understanding by all parties involved. Establishing prior informed consent also presumes that all directly affected communities will be provided complete information in an understandable form regarding the purpose and nature of the proposed programme, project, study or activities, the probable results and implications, including all reasonably foreseeable benefits and risks of harm (be they tangible or intangible) to the affected communities. Indigenous peoples, traditional societies and local communities have the right to make decisions on any programme, project, study or activities that directly affect them. In cases where the intentions of proposed research or related activities are not consistent with the interests of these peoples, societies or communities, they have a right to say no."

9. Principle of Confidentiality: "This principle recognises that Indigenous peoples, traditional societies and local communities, at their sole discretion, have the right to exclude from publication and/or to have kept confidential any information concerning their culture, identity, language, traditions, mythologies, spiritual beliefs or genomics. Parties to the research have a responsibility to be aware

of and comply with local systems for management of knowledge and local innovation, especially as related to sacred and secret knowledge. Furthermore, such confidentiality shall be guaranteed by researchers and other potential users. Indigenous peoples, traditional societies and local communities also have the rights to privacy and anonymity, at their discretion."

10. Principle of Respect: "This principle recognises the necessity for researchers to respect the integrity, morality and spirituality of the culture, traditions and relationships of Indigenous peoples, traditional societies, and local communities with their worlds."

11. Principle of Active Protection: "This principle recognises the importance of researchers taking active measures to protect and to enhance the relationships of Indigenous peoples, traditional societies and local communities with their environment and thereby promote the maintenance of cultural and biological diversity."

12. Principle of Precaution: "This principle acknowledges the complexity of interactions between cultural and biological communities, and thus the inherent uncertainty of effects due to ethnobiological and other research. The precautionary principle advocates taking proactive, anticipatory action to identify and to prevent biological or cultural harms resulting from research activities or outcomes, even if cause-and-effect relationships have not yet been scientifically proven. The prediction and assessment of such biological and cultural harms must include local criteria and indicators, thus must fully involve indigenous peoples, traditional societies, and local communities. This also includes a responsibility to avoid the imposition of external or foreign conceptions and standards."

13. Principle of Reciprocity, Mutual Benefit, and Equitable Sharing: "This principle recognises that Indigenous peoples, traditional societies, and local communities are entitled to share in and benefit from tangible and intangible processes, results and outcomes that accrue directly or indirectly and over the shorter and longer term for ethnobiological research and related activities that involve their knowledge and resources. Mutual benefit and equitable sharing will occur in ways that are culturally appropriate and consistent with the wishes of the community involved."

14. Principle of Supporting Indigenous Research: "This principle recognizes and supports the efforts of Indigenous peoples, traditional societies, and local communities in undertaking their own research

based on their own epistemologies and methodologies, in creating their own knowledge-sharing mechanisms, and in utilising their own collections and databases in accordance with their self-defined needs. Capacity-building, training exchanges and technology transfer for communities and local institutions to enable these activities should be included in research, development and co-management activities to the greatest extent possible."

15. Principle of The Dynamic Interactive Cycle: "This principle recognises that research and related activities should not be initiated unless there is reasonable assurance that all stages can be completed from (a) preparation and evaluation, to (b) full implementation, to (c) evaluation, dissemination and return of results to the communities in comprehensible and locally appropriate forms, to (d) training and education as an integral part of the project, including practical application of results. Thus, all projects must be seen as cycles of continuous and on-going communication and interaction."

16. Principle of Remedial Action: "This principle recognises that every effort will be made to avoid any adverse consequences to Indigenous peoples, traditional societies, and local communities from research and related activities and outcomes. Not withstanding the application of standards set out by this Code of Ethics, should any such adverse consequence occur, discussion will be had with the local peoples or community concerned to decide on what remedial action may be necessary to redress or mitigate adverse consequences. Any such remedial action may include restitution, where appropriate and agreed."

17. Principle of Acknowledgement and Due Credit: "This principle recognises that Indigenous peoples, traditional societies and local communities must be acknowledged in accordance with their preference and given due credit in all agreed publications and other forms of dissemination for their tangible and intangible contributions to research activities. Co-authorship should be considered when appropriate. Acknowledgement and due credit to Indigenous peoples, traditional societies and local communities extend equally to secondary or downstream uses and applications and researchers will act in good faith to ensure the connections to original sources of knowledge and resources are maintained in the public record."

18. Principle of Diligence: "This principle recognises that researchers are expected to have a working

> understanding of the local context prior to entering into research relationships with a community. This understanding includes knowledge of and willingness to comply with local governance systems, cultural laws and protocols, social customs and etiquette. Researchers are expected to conduct research in the local language to the degree possible, which may involve language fluency or employment of interpreters." (International Society of Ethnobiology, 2006 with 2008 additions)

I first coded each sentence from a random 15 of the 30 guidelines and licenses using the principles from the ISE Code of Ethics as published. I also added a *None* label for sentences that did not match any of the ISE Code of Ethics principles. For this first round of annotation, I noted places where the ISE Code of Ethics principles did and did not fit, especially for the licenses and for the signed language documents. Despite the emphasis on collaboration, the ISE Code of Ethics primarily addresses an academic audience and does not address the community members themselves or researchers who are themselves part of the community. The ISE Code of Ethics is also only concerned with indigenous community collaboration. My coding manual needed to include the perspectives of signed language communities, and address communities directly. These identified needs and the patterns observed in the first 15 guidelines and licenses supported changes to adapt the ISE Code of Ethics into my coding manual for the retrospective technical investigation. I found support across the guidelines and licenses for at least some aspects of each of the original ISE Code of Ethics principles. For some of the labels in the coding manual, I did not make any changes and for one I changed only the name. Several of the labels required substantial changes to both the name and the description of the label.

As discussed in Chapter 4, part of the challenge of adapting the ISE Code of Ethics principles to a coding manual is that the coding manual labels should be contrastive to ensure annotations made using the coding manual are systematic and reproducible. However, the ISE Code of Ethics principles are grounded in constellations of values which are interconnected and interdependent. Any categorization attempting to establish strict boundaries between the values will therefore be artificial and arbitrary to some extent. This means that the coding manual results were highly dependent on whichever 15 documents I first analyzed to adapt the ISE Code of Ethics principles. The coding manual should therefore not be interpreted as the ground truth of values related to collaborative community research, but rather a necessary instance of categorization for the purposes of conducting the analysis in this retrospective technical investigation.

The coding manual is presented in §5.3.1. I annotated the entire set of documents (30 guidelines and licenses and five archive policy documents) using this tuned coding manual. The results of that annotation process are detailed in §5.3.2. A second annotator, Dr. Emily M. Bender, then independently annotated 10% of the documents. In §5.3.3, I discuss the inter-annotator agreement as well as the implications of areas of disagreement. The trends in the coded documents and the trends in the annotation agreement contribute to the lessons learned from this retrospective exercise (§5.4).

5.3.1 Changes to the Coding Manual

As I began the annotation, it became clear that the ISE Code of Ethics would need general changes to address three mismatches. The first was that the ISE Code of Ethics was not developed as a coding manual. The second was a mismatch in the research focus. The International Society of Ethnobiology is focused on ethnobiology research and includes references to traditional knowledge as it applies to ethnobiology. The third mismatch is the target audience. The ISE Code of Ethics is addressed to its member researchers. It is not expected to apply to the community member collaborators, but it also does not consider researchers who themselves are part of the community. For this annotation project, a general research focus was more appropriate for spanning the diverse fields of research that the guidelines and licenses addressed and the topics should cover both researchers' and community collaborators' concerns.

Both the ISE Code of Ethics and the guidelines and licenses are framed as normative directives to the audience. These actions may explicitly state one or more values that these actions support, or the values may be implicit. Connecting the actions to the values requires interpretation, which is always dependent on the interpreter's own understandings of the context and values in mind. To avoid misattributing values to communities who do not state their values explicitly, I kept the coding manual in the directive style and directly incorporated recommended actions from the community guidelines and licenses. An alternative coding manual could have instead been reframed to code values rather than directives, starting from those values explicitly stated in both the ISE Code of Ethics text and the guidelines and licenses. However, this method would have either missed the implicit values in many directives or risked misattributing values to communities based on my own interpretations of possible values and the exact text. To limit the influence of my interpretations of values to the analysis stage of the investigation, I maintained the directive style in

the coding manual.

The changes to the coding manual resulted in reframing five labels with name changes, editing the text of eleven labels, and adding two labels. I left six labels in the coding manual unchanged (*Diligence*, *Confidentiality*, *Mindfulness*, *Remedial Action*, *The Dynamic Interactive Cycle*, and *Full Disclosure*). The full text of the final revised coding manual is presented in Appendix C.

Acknowledgement and Due Credit This label supports the recognition of the contributions to the research made by the research team and by the community by giving "due credit in all agreed publications and other forms of dissemination for their tangible and intangible contributions to research activities". Documents **NA4**, **NA3**, **SA3**, **OC5**, and **SL2** all provide recommendations along these lines or follow the recommendation themselves in the cases of **AS4**, **SA2**, and **SA4** (all licenses). The label also addresses the responsibility that researchers have in supporting that recognition in secondary or downstream uses and applications, placing on them a duty "to ensure the connections to original sources of knowledge and resources are maintained in the public record". This transfer of responsibility with the reuse of data was reflected in **SA2**, which instructed the users of the data to acknowledge the community in addition to the research team. The original text of the ISE Code of Ethics addressed to researchers, however, left the decision of whether or not to include community collaborators as co-authors up to the researcher. **SL2** and **NA4** suggest co-authorship alternatives, with **NA4** emphasizing that "A discussion with community collaborators about who should be acknowledged and how this should be done appropriately is recommended." I reframed this suggestion instead to "Co-authorship should be discussed with all contributors" to include the notion that the community collaborators may decide whether or not to be co-authors, as well as whether or not co-authorship is the best form of acknowledgement for each collaborator.

Active Protection The text from the ISE Code of Ethics regarding active protection focused on researchers as protectors responsible for "active measures to protect and to enhance the relationships of Indigenous peoples, traditional societies and local communities with their environment and thereby promote the maintenance of cultural and biological diversity". **AF3**, **SA3**, and **OC5** provide support for this recommendation. **OC5** argued that "researchers are expected to protect the rights and interests of Māori although there is little real involvement in the research process or outcomes." This emphasizes researchers' role in protecting

communities, but not the roles that communities may have in protecting their own community members and their traditional knowledge. In describing the establishment of the Inter-community Agreement for Equitable Access and Benefit Sharing between several indigenous communities in Peru, **SA3** states that the agreement also provides "a mechanism to protect and preserve traditional knowledge associated with biological resources and to strengthen the cultural identity of the communities." This example points to the roles that community leaders and researchers have in protecting their own cultural heritage and their communities. To shift the perspective of the coding manual text away from researchers as the sole protectors of communities, I added a new starting sentence broadening the possible categories of protectors and categories of protection recipients: "This principle recognises the importance of both community leaders, community researchers, and outside researchers taking active measures to protect indigenous peoples, traditional societies and local communities' traditional knowledge and their rights with respect to that knowledge." I then changed the original, now following, sentence to center relationship-enhancing research and traditional knowledge rather than the environment: "This principle encourages research that enhances the relationships of indigenous peoples, traditional societies and local communities with their traditional knowledge and thereby promotes the maintenance of cultural and biological diversity."

Collaboration and Active Participation Facilitating active participation and collaboration between indigenous communities and research teams in all phases of the research activities was a primary goal motivating the development of the ISE Code of Ethics. This included "collaboration on research design to address local needs and priorities, and prior review of results before publication or dissemination to ensure accuracy of information and adherence to the standards represented by this Code of Ethics". Many of the documents provided examples of community participation and collaboration, especially the community licenses such as **AF4**, **AS2**, **AS4**, **EU5**, **NA1**, **OC1**, and **SA2**. Some of the documents, however, advocated for collaboration beyond participation. Participation only requires that communities consent to and engage with the research as designed and carried out by an outside research team. Collaboration, however, requires that community members are also equal members of the directing research team and actively contribute to deciding the research goals and methods. **NA2** for example emphasizes that collaborative research ensures that "each aspect and phase of the project, from research design, to research team membership, to data gathering (interviews, surveys, observations, archival), to data interpretation, to forms of tribal review and to forms

of dissemination, are coordinated and implemented collaboratively, from co-designing and co-authorship." To reflect the priority of community collaboration over participation in research, I changed the name of this label from *Active Participation* to *Collaboration and Active Participation* and split the initial sentence to emphasize collaboration: "This principle recognises the crucial importance of collaboration between Indigenous peoples, traditional societies and local communities and research partners." I then reframed the reference to community members from research participants to collaborators and community leaders: "Community collaborators and leaders should actively contribute in all phases of research and related activities from inception to completion, as well as in application of research results."

Collective and Individual Inalienability Communities' inalienable rights to their cultural practices, traditional knowledge, and their languages were more often assumed than explicitly stated. **AF3** makes the inalienability of collective right clear: "The collective rights of indigenous peoples include recognition of their distinctive histories, languages, identities and cultures and the collective right to lands, territories and natural resources they have traditionally occupied and used, as well as the right to their collectively held traditional knowledge." **AF3** also shares the care that must be taken to uphold both the individual rights of community members and the collective rights of the entire community in the context of the commercialization of traditional knowledge: "In so far as it does not conflict with human rights violations of the individual, an individual cannot make this claim as one person. It is a community's collective claim in the context of the Khoikhoi community's customary laws." Individuals wanting to benefit from the commercialization must both identify themselves as community members and be identified as community members by other members. The text from the ISE Code of Ethics points out that the rights of communities to their traditional knowledge are "collective by nature but can include individual rights" and states, "It shall be for Indigenous peoples, traditional societies and local communities to determine for themselves the nature, scope and alienability of their respective resource rights regimes". Providing more specific actions in support of inalienability depends entirely on the collective rights and processes of the community in the collaboration at hand, so I did not change the text, however, I renamed this label from *Inalienability* to *Collective and Individual Inalienability* to highlight that both are equally important to this label.

Confidentiality Both the ISE Code of Ethics text and the documents described actions to support confidentiality for community members and for traditional knowledge. The text of the ISE Code of Ethics states, "This principle recognises that Indigenous peoples, traditional societies and local communities, at their sole discretion, have the right to exclude from publication and/or to have kept confidential any information concerning their culture, identity, language, traditions, mythologies, spiritual beliefs or genomics." This sentiment was supported by **NA4**, who advised prior agreement with communities on "the types of information and data that may be considered sensitive, and clarification on terms for confidentiality and data storage or disposal" with special considerations for "First Nations cultural knowledge and heritage, including traditional songs, stories, prayers, ceremonies, religious practices, rituals, plant or animal uses, techniques, designs, associated images, philosophies, and beliefs." The ISE Code of Ethics further states that, "Parties to the research have a responsibility to be aware of and comply with local systems for management of knowledge and local innovation, especially as related to sacred and secret knowledge." **OC5** points to this responsibility and considerations for safety being a minimum standard, with best practices advocating for directly incorporating Māori principles for trusting relationships and māhaki (respectful conduct) into the research. The transfer of this responsibility to maintain the confidentiality of the research participants and the traditional knowledge also extends to the users of the data, according to the ISE Code of Ethics. While this notion of transferred responsibility was not reflected in the initial 15 documents I coded, it was reflected in the archives and in several of the other 15 documents. For example, as part of the AILLA code of conduct, users agree that, "If the metadata for a resource states that names of creators and participants must be kept anonymous, I will respect their anonymity in any spoken or written representation of that resource that I produce." **OC1** also advises users of the data that, if somehow the anonymization procedures failed, the users would still be responsible for protecting the participants' confidentiality and "they should never link any particular still, clip or written annotated utterance, in a presentation or publication, with the name of the person or any other identifying information." Finally the ISE Code of Ethics text ends with the rights of individuals to privacy and anonymity. The documents from signed language communities particularly emphasized confidentiality for a numbers of reasons. **SL2** points to signed language communities' "close-knit nature and implications for confidentiality or anonymity in research"; **EU3** instantiates this individual confidentiality by anonymizing both the video data and the annotations for the data for any person names. **SL3**

also highlights the visual nature of signed languages as a consideration for confidentiality: "Sign linguists, whether Deaf or hearing, must also take into account the extra unavoidable dimension of video recording that makes it impossible to detach the data from the personal identity of the research participant. The rights and wishes of signers with regards to such recordings must be respected at all times." No changes were made to the ISE Code of Ethics text in the coding manual.

Diligence The core idea of diligence was supported by many documents, though the direct actions advocated for varied by community. The ISE Code of Ethics defined diligence as the recognition that "researchers are expected to have a working understanding of the local context prior to entering into research relationships with a community," including, "knowledge of and willingness to comply with local governance systems, cultural laws and protocols, social customs and etiquette." For **NA3** this means having an understanding of the historic trauma and its impacts on the present experiences of indigenous communities. For other communities such as **AF1**, **AF2**, and **AS3**, this means understanding the current social and economic forces affecting language change within their communities. Diligence also most directly engaged with the concept of language out of all the ISE Code of Ethics principles, stating that, "researchers are expected to conduct research in the local language to the degree possible, which may involve language fluency or employment of interpreters." Signed language communities especially emphasized the need for research conducted in signed languages and communication with signed language communities in their signed languages. **SL3** states that translators should be used "as a last resort" while **SL1** emphasizes that "the most accessible language for deaf people is the local sign language." This is because translators and interpreters for signed languages are usually hearing people and therefore are not the primary representatives of the community. For indigenous communities however, translators and interpreters are often indigenous people themselves who have a central role in analyzing data and research results. For this reason, I chose to keep the language in the ISE Code of Ethics that first emphasizes that research should be conducted in the community's language and secondarily offers possible means to achieve this goal, assuming that those means will be negotiated with the community. Therefore, no changes were made to the ISE Code of Ethics text in the coding manual.

Diversity and Representation Several of the documents described the need for recognition of the internal diversity of communities. Some referred to this diversity in the context of collective rights to traditional

knowledge and decision-making. For example, **SA3** describes the Inter-Community Agreement supporting collective decision-making through leaders of the six communities as well as a general assembly of elected representatives. Others referred to diversity in the context of representation within the research and disseminated research results. This consideration for diversity is relevant for defining research outputs (such as second language materials for diaspora communities wanting to learn the language in **AF4**), for who is providing community perspectives to the research (e.g., **SL1** and **SL3**), and how communities may change over time and be in the process of changes (e.g. **SA1**, **SL2**, and **AF1**). To reflect these considerations in the coding manual, I drew from **SL2**: "investigators should recognize the diverse experiences, understandings, and way of life (in sign language societies) that reflect their contemporary cultures." Keeping with the directive style of the coding manual, I added the label *Diversity and Representation*, starting with: "This principle recognizes the diverse experiences, understandings, and way of life that reflect Indigenous peoples, traditional societies and local communities' contemporary cultures." The text also addresses the respect for community changes over time and resulting diversity within the community. The final sentence maintains the importance of community-led demographic categories in both research and governance: "It is up to the community to determine what this diversity looks like and what appropriate representation within the community looks like for both descriptive characterizations and decision-making bodies."

Free Prior Informed Educated Consent Consent is a process embedded in many of the documents, usually in reference to the United Nations Declaration on the Rights of Indigenous Peoples, which states that indigenous peoples have the right to free, prior, and informed consent (FPIC) in Articles 10, 11, 19, 28, 29, and 32 (Office of the High Commissioner for Human Rights, 2007). **AF3** emphasized the requirements for free consent "given voluntarily and without coercion, intimidation or manipulation," namely that community members give their consent "in an environment where they do not feel intimidated, and where they have sufficient time to discuss in their own language, and in a culturally appropriate way." To reflect this consideration, I changed the name of this label from *Educated Prior Informed* to *Free Prior Informed Educated Consent* and added "free" prior to consent within the text of the coding manual. Although "informed" and "educated" are similar, "educated" connotes more effort towards ensuring that the party who is being asked for their consent understands the terms they are consenting to, so I kept it in the label but opted to maintain the order of the adjectives in line with FPIC as used in Office of the High Commissioner for Human Rights

(2007) and many of the documents. The remainder of the text from the ISE Code of Ethics was supported by the selected documents: consent must be obtained prior to the research starting (e.g., **AF3**); it may involve collective processes (e.g., **OC5**); and it is an ongoing process throughout the research (e.g., **OC2**). The ISE Code of Ethics also points to community consent requiring "an educative process that employs bilingual and intercultural education methods and tools, as appropriate, to ensure understanding by all parties involved." **SL2** also advocates for consent methods that "make the consultant aware of all the implications of providing data, of being video recorded, and (when applicable) of the long-term archiving and sharing of the obtained data as well as of the implications of research itself," while **OC5** emphasizes "recognising the place of oral consent in some Māori settings." Further concerns around consent indicated in the ISE Code of Ethics include communicating the goals of the research and possible benefits and risks for the community (as also discussed in **EU4**) and the rights of communities to make decisions on any research that impacts those communities, including the right to say no to the research (as discussed in **AF5** and others).

Full Disclosure Similar to consent, full disclosure involves expectations around how researchers (including community researchers) communicate with the community at large. The text of the ISE Code of Ethics states that, "This principle recognises that Indigenous peoples, traditional societies and local communities are entitled to be fully informed about the nature, scope and ultimate purpose of the proposed research (including objective, methodology, data collection, and the dissemination and application of results)." **OC2**, **NA4**, and **EU4** all provide similar statements around fully providing information to communities and describe the processes that communities have in place to support this communication. The ISE Code of Ethics also affirms that the information described in the prior sentence "is to be given in forms that are understood and useful at a local level and in a manner that takes into consideration the body of knowledge, cultural preferences and modes of transmission of these peoples and communities." Both **AF5** and **SL3** provide support for this. **AF5** explains that, "complex issues must be carefully and correctly described" without assuming that the community cannot understand. **SL3** similarly advocates for explanations provided in "a simple, accessible way" particularly for signed language communities where the most accessible form of communication is signed language. No changes were made to the ISE Code of Ethics text in the coding manual.

Mindfulness This concept was included in the ISE Code of Ethics preamble, rather than in the principles, and was included in the comparison exercise by Tunón et al. (2016). The ISE Code of Ethics defines "mindfulness" as "an obligation to be fully aware of one's knowing and unknowing, doing and undoing, action and inaction." While the term "mindfulness" was not explicitly used in the guidelines and licenses, discussions of positionality in research and awareness of one's actions in the context of entrenched systems of oppression and exploitation did occur in several of the analyzed documents. For example, **NA3** reminds its audience that, "Research is historically-situated, geo-politically positioned, relational, and explicit about the perspective from which knowledge is generated... There is always an historical context to Indigenous knowledge and praxis, which are inseparably linked to Indigenous identity and all its interrelated socio-political expressions." This connection between community histories and their impacts on current research was also brought up in the context of signed language communities by **SL3**: "Sign language users and communities have been traditionally marginalized and researchers must always be aware that this might result in power inequalities between sign language consultants and researchers." Somewhat surprisingly, **EU4** was one of the few documents to explicitly discuss the unique positionality of community researchers: "Sámi researchers doing research in their communities already have a standing as members of their communities, and thus the issues they need to consider are different in comparison to researchers coming from outside the communities." Elucidating the different considerations for community researchers was left for future work (see Chapter 9 for further discussion). No changes were made to the ISE Code of Ethics text in the coding manual.

Ownership and Permission Data ownership and permitted uses were the primary concerns of many of the licenses included in the selected documents. While the ISE Code of Ethics principles included prior rights and responsibilities in the context of rights to intellectual property and traditional resources (discussed further below), the focus of prior rights excludes considerations for future use and the nuances of ownership transfer. To address these considerations in the coding manual, I added the label *Ownership and Permission*. **SL2** points out that hegemonic forces in academic research have contributed to the belief that "the researchers—not the participants in the project—have ownership of the intellectual property." I reframed this point as a normative statement within the coding manual: "This principle recognizes that communities are the rightful owners of their communal knowledge." Some of the guidelines also pointed

to community processes for exercising control over data practices in research. For example, **NA4** states that, "the Nuu-chah-nulth Tribal Council is evoking specific restrictions about data ownership, storage and permission for access to all data collected in the Nuu-chah-nulth communities." Within the licenses, many contain restrictions on use and redistribution without written permission (e.g., **EU2** and **EU3**). The next sentence in the coding manual text again sets as a normative statement that researchers will respect these mechanisms by which communities maintain control of their data: "Researchers will follow the processes set by communities to ask for permission to engage with the communities and their knowledge practices and for any reuse or redistribution of community knowledge." Examples of permitted and disallowed uses that I saw in the labeled documents were added as the final sentence of the coding manual text: "Communities may define their own permitted uses (such as for research, education, or personal uses only) or may define disallowed uses (such as commercial)." Sometimes examples permitted only research uses (e.g., **EU1**), only education uses (e.g., **AS1**), and sometimes both (e.g., **OC3**). Permissions for commercialization uses also varied from being generally disallowed (e.g., **SA1**) to commercialization being allowed with permission and benefits returned to the community (e.g., **AF3**). **OC4**, a license applying to the Te Hiku Kaituhi automatic transcription tool[12], specifically disallows "any attempt to use Kaituhi to build Māori data sets and natural language processing models" to maintain community control of data and language technology development, which often includes commercialization. Permissions for use and the nuances of data ownership were also topics covered by all of the archives.

Precaution The original text of the ISE Code of Ethics was primarily concerned with biological harms from research, however, removing the text specifying biological harms rendered the text usable for considering general harms from research. For example, the first sentence of the principle in the ISE Code of Ethics is as follows: "This principle acknowledges the complexity of interactions between cultural and biological communities, and thus the inherent uncertainty of effects due to ethnobiological and other research." In the coding manual, removing the specific references to biological topics reframes the sentence to show concern for the impacts of general research and to motivate active efforts by the research team to prepare for both intended and unintended results: "This principle acknowledges the complexity of interactions, and thus the inherent uncertainty of effects due to research." **NA4** consider this uncertainty in the possible side-effects

[12] `https://kaituhi.nz/`

of the dissemination of research on cultural heritage, including attracting attention from a broad, unknown audience, thereby making the cultural knowledge "particularly susceptible to appropriation, exploitation and commodification, whether or not this is the intent of the researcher." The ISE Code of Ethics text "advocates taking proactive, anticipatory action to identify and to prevent harms resulting from research activities or outcomes." For **NA2**, these harms include violations of trust between the community and the researchers as a result of either intentional or unintentional actions, reinforcing that the impact on the community is the priority in considering precautionary actions. The ISE Code of Ethics text centers researchers and Western conceptions of burden of proof, i.e., "fully involving Indigenous peoples" in predicting and assessing harms and anticipating harms "even if cause and effect relationships have not yet been scientifically proven." For the coding manual, I removed the second phrase concerning proof and edited the phrasing on predicting and assessing harms from "involving" communities to "prioritizing the perspectives" of communities: "The prediction and assessment of such harms must include local criteria and indicators, thus must prioritize the voices of Indigenous peoples, traditional societies, and local communities." While transparent assessments of possible harms are supported by communities (e.g., **AF5**), suggesting that cause and effect relationships may need to be "scientifically proven" at some point diminishes communities' own assessments of the research impacts and their experiences during the research. I added additional considerations for possible research harms coming from a lack of commitment on the part of outside researchers in ensuring that communities also benefit from the research to the final sentence of the ISE Code of Ethics text: "This also includes a responsibility to avoid the imposition of external or foreign conceptions and standards and to commit to the project for the duration of its investigation and reintegration of resulting benefits to the community." **SL1** both provided the example for a lack of commitment to the research and the community contributing to the possible harms of research and supported the initial concern of imposing outside standards on communities, advising that researchers "avoid the assumption that deaf people and communities in the developing world have a desire for Western solutions, live in a context with high levels of resources, or that western ideas, concepts and developments can be 'copy-pasted' into foreign contexts."

Prior Rights and Responsibilities In discussing prior rights and responsibilities of indigenous communities, the ISE Code of Ethics text affirms that "Indigenous peoples, traditional societies, and local communities have prior, proprietary rights over, interests in and cultural responsibilities for all air, land, and

waterways, and the natural resources within them that these peoples have traditionally inhabited or used, together with all intellectual property and traditional resource rights associated with such resources, knowledge, and their use." For **NA2**, researcher recognition of prior rights "entails acknowledging the intellectual and cultural property rights, as well as Tribal sovereignty of the federally-recognized American Indian tribe with whom you are seeking to initiate a research project," and recognizing that "the tribal collaborators have as much responsibility and governance over products – reports, maps, datasets, etc. – as investigators." In the original text of this label, "knowledge" is included with recognition of intellectual property and traditional resource rights over resources and their use, but intellectual property and rights also apply to the knowledge itself, such as the research products described by **NA2**. To reflect this, I made a minor edit to the sentence to include the acknowledgement of knowledge with resources and their use rather than intellectual property and traditional resource rights. Licenses like **OC4** are one mechanism for communicating and enforcing rights; while other licenses base their text in language from the archives and from Creative Commons licenses, **OC4** was developed as "an international example for indigenous people's retention of mana over data and other intellectual property in a Western construct," where *mana* refers to the te reo Māori concept of justice and equity, reflected through power and authority (from **OC5**). **EU4** highlights the necessity of legal expertise in these matters to ensure adaptability to "differing national legislations and regulations, such as with matters of personal details, data protection, publicity and equality laws, as well as laws regulating the academia." As discussed in §2.2, establishing rights and recognition for signed languages has been difficult in many countries, so there were no examples from the selected signed language documents on this topic.

Reciprocity, Mutual Benefit, and Equitable Sharing Ensuring mutual benefits for communities in research contexts is the primary focus of two of the documents (**AF5** and **SA3**) and the ISE Code of Ethics text emphasizes the importance of ensuring mutual benefits in its first sentence: "This principle recognises that Indigenous peoples, traditional societies, and local communities are entitled to share in and benefit from tangible and intangible processes, results and outcomes that accrue directly or indirectly and over the shorter and longer term for ethnobiological research and related activities that involve their knowledge and resources." I removed the term "ethnobiological" in this sentence in the coding manual text, as it narrowed the scope of research fields which communities are entitled to benefit from. For **NA2** and others, mutual benefits must be "meaningful and applicable to your host community, as defined by your host the host com-

munity." The emphasis on defining benefits according to the community's terms points to instances of past research where no effort was made to engage with the needs of the community, and benefits were defined in terms of academic outputs such as publications. The second sentence of the ISE Code of Ethics text affirms that, "Mutual benefit and equitable sharing will occur in ways that are culturally appropriate and consistent with the wishes of the community involved." I also added a sentence encouraging research partners to "consider the benefits they receive from community collaborators and the lessons that may be brought back to the academic community." This sentence was inspired in part by **SL2**, referencing Lincoln and Denzin (2005): "Those who take such chances in research that departs from the conforming standards imposed by those who hold academic power in fact teach the latter a thing or two". **SA3** echoes this notion that community epistemologies can be integrated with Western ones "through sharing experiences, including experiences of overcoming obstacles, and ideas about best practice in the design and implementation of a participatory, creative methodology and framework for benefit sharing." Though inappropriate applications of these epistemologies may contribute to cultural appropriation, careful applications developed with communities can contribute to a broader awareness of other ways of knowing and equal consideration for those other ways of knowing as valid within academic communities.

Remedial Action Only a few of the documents mentioned remedial action, but those documents made it clear that communities hold researchers accountable for the impacts of their research. **OC5** points out that many of the guidelines themselves have been developed "in response to examples of research that resulted in adverse outcomes and/or experiences for participants and their communities." The ISE Code of Ethics text states that, "every effort will be made to avoid any adverse consequences to Indigenous peoples, traditional societies, and local communities from research and related activities and outcomes. Not withstanding the application of standards set out by this Code of Ethics, should any such adverse consequence occur, discussion will be had with the local peoples or community concerned to decide on what remedial action may be necessary to redress or mitigate adverse consequences. Any such remedial action may include restitution, where appropriate and agreed." This framing leaves the definition of adverse consequences and the appropriate remedial action up to the community. **AF5** references the "use of dispute resolution mechanisms" to enforce compliance with their own research code of ethics as well as consequences for the researchers, their institutions, and future collaborations. **OC2** lists possible consequences such as withdrawal of fund-

ing; written censure of researchers; suspension of research contracts; and community withdrawal from the research. No changes were made to the ISE Code of Ethics text in the coding manual.

Respectful Relationships The ISE Code of Ethics text frames respect in the following context: "This principle recognises the necessity for researchers to respect the integrity, morality and spirituality of the culture, traditions and relationships of Indigenous peoples, traditional societies, and local communities with their worlds." However, this unidirectional system of respect, from researchers to communities, misses many aspects of what communities define as part of respect. **OC5** advocates for "constructive relationships and acknowledges the roles, relationships and responsibilities each party has in the process of engagement." **NA3** also frames respect within research and research practices as bidirectional: "trauma-informed approaches begin with establishing trust, friendship, and mutual respect." To highlight the relationship aspects of respect, I renamed the label *Respectful Relationships* in the coding manual and added a new starting sentence: "This principle advocates for respectful, constructive relationships and acknowledges the roles, relationships and responsibilities each party has in the process of engagement." The original sentence from the ISE Code of Ethics then follows as secondary to the relationship-building efforts, but still necessary. **NA4** includes "patience, respect and appreciation for the people, creatures and places in whose communities you are a guest" as part of ensuring successful collaborations and building opportunities for future collaborations. However, **SL2** warns against hierarchical relationships in research, advocating instead for "a horizontal dialogue between research teams and participants." Considering this suggestion and related discussions on power dynamics in the context of systems of oppression, I added a final sentence to this label in the coding manual: "Research collaborations should also be aware of the power dynamics between outside researchers, community researchers, and community research participants and endeavor to approach hierarchical power structures in a culturally-appropriate manner." With this framing, hierarchical structures may still exist if they are appropriate for the community's self-defined social structure.

Self-determination The UNDRIP and several other international agreements assert the right to self-determination for all peoples (Office of the High Commissioner for Human Rights, 2007). The ISE Code of Ethics text for self-determination concurs with those agreements: "This principle recognises that Indigenous peoples, traditional societies and local communities have a right to self-determination (or local

determination for traditional and local communities) and that researchers and associated organisations will acknowledge and respect such rights in their dealings with local peoples and their communities." As discussed above, community self-determination and community definitions support many of the discussions around consent, benefits, remedial action, and others; **SA3** describes self-determination itself as "the right of people to control their own resources, economies and livelihoods, and to choose what cultural values they will embrace." **EU4** and **SL2** also highlight the importance of self-determination in research, with **SL2** stating that self-determination "empowers the community members to take a stand on how researchers may investigate them." From these considerations, I added a sentence to the coding manual: "This right extends to decisions made about research topics and appropriate investigation methods when the research concerns the community and their knowledge."

Supporting Community Research I changed the ISE Code of Ethics' label from *Supporting Indigenous Research* to *Supporting Community Research* to be more inclusive for both signed language communities and other communities who may not refer to themselves as indigenous (see Chapter 2 for discussion of the term *indigenous*). The ISE Code of Ethics text states, "This principle recognizes and supports the efforts of Indigenous peoples, traditional societies, and local communities in undertaking their own research based on their own epistemologies and methodologies, in creating their own knowledge-sharing mechanisms, and in utilising their own collections and databases in accordance with their self-defined needs." **AF3** added that research "should consider the community's own legitimate decision-making processes regarding all phases of planning, implementation, monitoring, assessment, evaluation and wind-up of a research project." While Western methods of inquiry may be integrated for some questions related to community research, **NA2** emphasizes that "if the goal is to appreciate, understand and apply the meanings, structures and dynamics of the Indigenous, to utilize a route to the summit other than an Indigenous methodology would only result in reaching a 'false summit,' and a distorted, if not misinterpreted view of the Indigenous." Truly understanding communities requires centering their definitions and ways of knowing. To support this idea, I added the following sentence to the coding manual: "Research with local communities should make efforts to center community perspectives and conduct such research in culturally relevant environments, rather than those that are convenient for outside researchers." In addition to centering community research methods, centering important places for research was also a topic addressed by **NA3**, who called for "the use of ef-

fective practices that generate concrete knowledge in interrelated and vibrant social environments; not just in environments that are efficient in data gathering." The ISE Code of Ethics text also endorses supporting communities in constructing and sustaining their own research infrastructures: "Capacity-building, training exchanges and technology transfer for communities and local institutions to enable these activities should be included in research, development and co-management activities to the greatest extent possible" Capacity building for communities was one of the main focuses of **SL1** and was exemplified in licenses like **AS2** and **OC4**.

The Dynamic Interactive Cycle The ISE Code of Ethics text for the dynamic interactive cycles of research states, "This principle recognises that research and related activities should not be initiated unless there is reasonable assurance that all stages can be completed from (a) preparation and evaluation, to (b) full implementation, to (c) evaluation, dissemination and return of results to the communities in comprehensible and locally appropriate forms, to (d) training and education as an integral part of the project, including practical application of results." As stated simply by **AF5**, "Respect requires that promises made by researchers need to be met." While challenges for the research may require the research team to change their initial methods or scope, **NA4** encourages researchers to communicate these changes and their rationales to the community "with regular written or verbal updates on your research progress, your current contact information, and opportunities for feedback." **SL1** advises that maintaining the original goal of the research is best done as "part of a long-term plan with and for the community." Dissemination of the results or the data may be done best in the community's language, such as **EU1** who provided a community license in German, and tailored to address community-specific concerns, like **OC5** who encourage researchers to communicate results "focused on matters of relevance to Māori with information directed to an end use that shows clear benefits for Māori." The ISE Code of Ethics text further notes that, "Thus, all projects must be seen as cycles of continuous and on-going communication and interaction." This metaphor aligns with communities' advice and own research methods. **NA2** advises research teams initiated by organizations outside the community to continue community consultation for a research project "during the research phase through to its conclusion and dissemination of the resulting research." As a result of communities' own internal changes and patterns of language change over time, **OC1** states that its own dataset, Auslan Signbank, "is constantly changing and being updated with new information contributed by the deaf community or derived

from annotations in the Auslan Corpus." Ongoing engagement with the community ensures that the project stays relevant and up-to-date with community language practices. Communication between communities and broader audiences was also discussed in the documents. **SA3**, in explaining the reasoning for putting in writing a community agreement based in oral customary laws, lists supporting other communities in establishing their own similar agreements, communication to broader research fields, and providing an example of the practical applications of such an agreement as motivation for deviating from traditional communication methods. Such examples highlight that not only is communication important between the community and the research team, but also between communities and allied research teams around the world. No changes were made to the ISE Code of Ethics text in the coding manual.

Traditional Guardianship This label concerns actions that recognize and support communities in protecting their traditional knowledge and ensuring that knowledge is transmitted to the next generation. **NA2** defines traditional knowledge as information that is "generated, preserved and transmitted in a traditional and intergenerational context; distinctively associated with a tribe which preserves and transmits it between generations; integral to the cultural identity of the tribe, which holds the knowledge through a form of custodianship, guardianship, collective ownership, or cultural responsibility." The ISE Code of Ethics text states, "This principle recognises the holistic interconnectedness of humanity with the ecosystems of our Sacred Earth and the obligation and responsibility of Indigenous peoples, traditional societies and local communities to preserve and maintain their role as traditional guardians of these ecosystems through the maintenance of their cultures, identities, languages, mythologies, spiritual beliefs, and customary laws and practices, according to the right of self-determination." **EU4** asserts traditional guardianship as indigenous peoples' "right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions, as well as the manifestations of their sciences, technologies and cultures." **AS2**, in documenting the Hrusso Aka community's traditional knowledge, notes that, "The stories, mythological narratives and oral histories found here are the versions of their narrators and we do not claim that they are the only or the most accurate versions." To reflect this notion of respect for diversity within a community's traditional knowledge, I added "internal diversity" to the list of concepts within the purview of traditional guardianship. Whereas the Diversity and Representation label focuses on the diversity of the people in the community, this aspect of diversity is intended to address the diversity within a community's traditional

knowledge, such as the Hrusso Aka community's variations in their narratives. While the ISE framing of traditional guardianship focuses more on ecological guardianship, **SA3** ties together indigenous identity and biological diversity, arguing that indigenous governance of the land under the biocultural protocol of the local community "gives a holistic value to indigenous territoriality (not its commodification), re-establishing and enhancing old and new biocultural networks." Enhancing indigenous peoples' abilities to define their relationships with their ancestral territories therefore supports their cultural identity and their community self-determination. Traditional guardianship still applies to signed languages communities, who typically do not associate their cultural identity with territorial rights. For signed language communities, the focus is on culture, language, and the experience of living as a deaf person in a hearing society; **SL1** therefore defines deaf signers as "the real experts on language issues." **NA3** also places an emphasis on indigenous peoples' "lived reality that does not need to be mediated, translated, and interpreted to gain mainstream academic legitimacy." To support this notion of expertise through lived experience (borrowing from Young et al. (2019)'s concept of *experiential experts*), I added the following sentence to the coding manual: "This principle also recognizes that communities are experiential and cultural experts of their traditional knowledge, and this knowledge is valid and legitimate without further mediation, translation, or interpretation." With this addition, traditional guardianship is not only a responsibility, but also a position of authority as communities know how best to protect traditional knowledge.

Summary I found support for all of the labels derived from the ISE Code of Ethics within the first 15 documents I coded, despite some changes to the text of the coding manual. With evidence from the licenses, I added a new *Ownership and Permission* label to address third party uses of data and the mechanisms communities use to maintain control of their data. I also added a new *Diversity and Representation* label to reflect communities' emphasis on community-defined diversity and accurate representation within research and within published descriptions of the community. Table 5.9 summarizes the types of changes that I made to adapt the ISE Code of Ethics principles into coding manual labels. With the coding manual finalized, I coded all of the 35 documents using the coding manual labels, including the original 15 documents I analyzed for creating the coding manual.

Type of Changes	# of Labels	Labels
No Change	6	Confidentiality Diligence Full Disclosure Mindfulness Remedial Action The Dynamic Interactive Cycle
Name Only	1	Collective and Individual Inalienability
Description Only	7	Acknowledgement and Due Credit Active Protection Precaution Prior Rights and Responsibilities Reciprocity, Mutual Benefit, and Equitable Sharing Self-determination Traditional Guardianship
Name and Description	4	Collaboration and Active Participation Free Prior Informed Educated Consent Respectful Relationships Supporting Community Research
New	2	Diversity and Representation Ownership and Permission

Table 5.9: Summary of changes to adapt the ISE Code of Ethics principles into coding manual labels.

5.3.2 Coding Results

I annotated a total of 3355 sentences across all the documents, including the policy documents from the archives. Table 5.10 shows the percentages of sentences from each region that were annotated with each label as well as the percentages of all sentences that were annotated with each label. By far the most used label was *None* at 38.8% of the labels. The next most common label was *Supporting Community Research* at 5.9% of the labels. *Remedial Action* was the least used label at 0.7% of all labels. Text that was labeled as *None* included: structural text such as headings, titles, and the beginning text of lists; legal text in the selected documents not tied to community values; background information on the community context, academic research methodologies (as opposed to community methodologies), and related or example research projects; statements on the contents of archives and their collection and annotation processes (when not tied to community values, as seen in the licenses authored by non-community affiliated research teams); and references to other guidelines. As *None* was the most common label across all the categories, it is excluded from the discussions and tables in the following analysis.

Supporting Community Research, *Diligence*, and *Reciprocity, Mutual Benefit, Equitable Sharing* were common across all of the categories, but which label was the most common varied by geographic region as shown in Table 5.11. The most used labels across the documents from Africa were *Diligence*, then *Reciprocity, Mutual Benefit, Equitable Sharing*, then *Traditional Guardianship*. For the documents from Asia, the most common labels were *Acknowledgement and Due Credit*, *Diligence*, and *Diversity and Representation*. The most used labels for the documents from North America were *Supporting Community Research*, *Collaboration and Active Participation*, and *The Dynamic Interactive Cycle*. Unlike Africa, Asia, and North America, *Prior Rights and Responsibilities* was much more likely to be used for documents from Oceania, South America, and Europe. For South America, is was the most commonly used label, followed by *Reciprocity, Mutual Benefit, Equitable Sharing* and *Supporting Community Research*. The documents from Oceania were also likely to be labeled with *Reciprocity, Mutual Benefit, Equitable Sharing* and *Prior Rights and Responsibilities*, followed by *Ownership and Permission*. The documents from Europe were the only group in which none of *Supporting Community Research*, *Diligence*, and *Reciprocity, Mutual Benefit, Equitable Sharing* were in the top three most used labels. Instead, *Prior Rights and Responsibilities*, *Ownership and Permission*, and *The Dynamic Interactive Cycle* were the most common.

Label	Africa	Asia	Europe	North America	Oceania	South America	All Regions
A&DC	3.46	11.55	2.30	3.32	3.08	4.81	4.02
AP	1.15	1.44	3.26	1.66	3.33	2.61	2.35
C&AP	1.38	6.14	3.07	8.30	3.72	3.41	4.62
C&II	3.69	0	0.96	0.47	0.51	1.60	1.10
C	0.46	0.36	1.34	1.90	1.28	0.20	1.10
D	8.53	9.03	1.34	4.98	2.44	4.81	4.59
D&R	3.00	6.86	2.49	2.25	0.90	7.21	3.19
FPIEC	4.84	2.53	2.11	2.61	2.69	0.20	2.47
FD	3.69	0.72	0.96	2.73	1.28	0.60	1.76
M	0	0.36	0.57	2.02	0.38	0.60	0.80
N	39.63	49.10	56.13	25.98	40.90	32.87	38.84
O&P	1.84	2.17	4.60	3.20	7.56	2.40	4.05
P	0.69	0.72	2.49	1.90	2.18	2.40	1.88
PR&R	5.07	1.44	6.13	3.56	7.56	9.22	5.75
RMBES	6.68	2.53	0.57	4.15	7.82	8.42	5.28
RA	0.92	0	0.96	0.36	1.54	0	0.72
RR	2.53	0	2.49	4.03	2.18	0.40	2.30
SD	2.76	0.72	1.15	2.37	1.15	1.60	1.70
SCR	2.53	2.17	1.34	12.69	3.97	7.41	5.93
TDIC	1.61	1.44	4.41	6.76	4.10	3.21	4.14
TG	5.53	0.72	1.34	4.74	1.41	6.01	3.40
Sum	100	100	100	100	100	100	100

Table 5.10: Percentage of sentences from each region that were annotated with each label.

Region	Label	Count	% of Region Total
Africa	Diligence	37	8.5%
	Reciprocity, Mutual Benefits, Equitable Sharing	29	6.7%
	Traditional Guardianship	24	5.5%
Asia	Acknowledgement and Due Credit	32	11.6%
	Diligence	25	9.0%
	Diversity and Representation	19	6.9%
Europe	Prior Rights and Responsibilities	32	6.1%
	Ownership and Permission	24	4.6%
	The Dynamic Interactive Cycle	23	4.4%
North America	Supporting Community Research	107	12.7%
	Collaboration and Active Participation	70	8.3%
	The Dynamic Interactive Cycle	57	6.8%
Oceania	Reciprocity, Mutual Benefits, Equitable Sharing	61	7.8%
	Prior Rights and Responsibilities	59	7.6%
	Ownership and Permission	59	7.6%
South America	Prior Rights and Responsibilities	46	9.2%
	Reciprocity, Mutual Benefits, Equitable Sharing	42	8.4%
	Supporting Community Research	37	7.4%
All Regions	Supporting Community Research	199	5.9%
	Prior Rights and Responsibilities	193	5.8%
	Reciprocity, Mutual Benefits, Equitable Sharing	177	5.3%

Table 5.11: Top three most frequent labels by geographic region (excluding *None*). Percentage calculated using total number of annotated sentences for each region.

Modality	Label	Count	% of Modality Total
Signed	Diligence	51	12.1%
	Acknowledgement and Due Credit	37	8.8%
	Diversity and Representation	34	8.1%
Spoken	Supporting Community Research	166	6.6%
	Reciprocity, Mutual Benefits, Equitable Sharing	161	6.4%
	Prior Rights and Responsibilities	158	6.3%

Table 5.12: Top three most frequent labels by modality (excluding *None*). Percentage calculated using total number of annotated sentences for each modality.

Author type	Label	Count	% of Author Type Total
Community	Reciprocity, Mutual Benefits, Equitable Sharing	159	6.2%
	Supporting Community Research	153	6.0%
	Prior Rights and Responsibilities	149	5.8%
R/Gov	Supporting Community Research	46	12.4%
	Prior Rights and Responsibilities	44	10.7%
	Ownership and Permission	37	9.0%

Table 5.13: Top three most frequent labels by author type (excluding *None*). Percentage calculated using total number of annotated sentences for each author type.

Looking at the labels across modalities in Table 5.12, the documents representing spoken languages align with the patterns of the geographic regions (since the spoken languages are a larger subset of the geographic regions). For the spoken modality, *Supporting Community Research*, *Reciprocity, Mutual Benefit, Equitable Sharing*, and *Prior Rights and Responsibilities* are the most common non-*None* labels. For signed languages, however, the most common labels are *Diligence*, *Acknowledgement and Due Credit*, and *Diversity and Representation*. The prevalence of the *Diligence* label in signed language documents follows from the logical, yet still needed, insistence that researchers conducting research on signed language topics have at least some proficiency in the signed language of interest. While it is possible to include interpreters in good research projects, it is the researcher's responsibility that the skills of the interpreters are an appropriate match for the research and the community, which still requires proficient knowledge of the signed language of interest. Many signed languages are still working to establish legal recognition in the countries where they are used (De Meulder et al., 2019b), which may have contributed to the reduced use of the *Prior Rights and Responsibilities* label.

The community authored documents follow the primary trend of *Reciprocity, Mutual Benefit, Equitable Sharing*, *Supporting Community Research*, and *Prior Rights and Responsibilities* occurring most frequently. Across all of the documents authored by research and government organizations (including the archive documents), the most common labels are *Supporting Community Research*, *Prior Rights and Responsibilities*, and *Ownership and Permission*. *Reciprocity, Mutual Benefit, Equitable Sharing* is the 10th most common label among research and government-authored documents. Including more suggestions for reciprocity in these documents could support the development of research projects more closely aligned with community

Document type	Label	Count	% of Document Type Total
Licenses	Acknowledgement and Due Credit	105	10.6%
	Ownership and Permission	71	7.2%
	Prior Rights and Responsibilities	67	6.8%
Guidelines	Supporting Community Research	182	9.4%
	Reciprocity, Mutual Benefit, Equitable Sharing	142	7.3%
	Collaboration and Active Participation	121	6.2%
Archives	Prior Rights and Responsibilities	34	8.0%
	Ownership and Permission	29	6.8%
	Active Protection	26	6.1%

Table 5.14: Top three most frequent labels by document type (excluding *None*). Percentage calculated using total number of annotated sentences for each document type.

goals.

There were significant differences between the most common labels across the document types. For licenses, perhaps unsurprisingly, the main labels were *Acknowledgement and Due Credit*, *Ownership and Permission*, and *Prior Rights and Responsibilities*. For guidelines, again following the majority trend, the most frequent labels were *Supporting Community Research*, *Reciprocity, Mutual Benefit, Equitable Sharing*, and *Collaboration and Active Participation*. The archive documents were the only category to have *Active Protection* among the top labels, along with *Prior Rights and Responsibilities* and *Ownership and Permission*. The archive documents were also the only category to have no instances of one or more labels. *Collective and Individual Benefit*, *Diversity and Representation*, and *Collaboration and Active Participation* were not used once across any of the archive documents. This marked difference in annotation results suggests that the archive documents could be revised to better align with community needs and goals.

5.3.3 Inter-Annotator Agreement

To determine the reliability of this coding protocol, a second annotator, Dr. Emily M. Bender, independently annotated four of the documents using the final coding manual (10% of the 35 documents, rounded up). These included documents from an archive (the license, terms of access, and conditions of use for AILLA), from a signed language perspective (World Federation of the Deaf Expert Group on Developing Countries, 2016), from a community perspective (Kimura (depositor), 2018; Ka Haka 'Ula O Ke'elikolani College of

Label	A&DC	AP	C&AP	C&II	C	D	D&R	FPIEC	FD	M	N
A&DC	2	-	-	-	-	-	-	-	-	-	1
AP	-	-	-	-	-	-	-	-	-	-	5
C&AP	-	-	3	-	-	-	-	-	-	-	-
C&II	-	-	-	-	-	-	-	-	-	-	-
C	-	-	-	-	1	-	-	-	-	-	-
D	-	1	-	-	-	8	-	-	-	1	1
D&R	-	-	2	-	-	-	1	-	-	-	-
FPIEC	-	-	-	-	-	-	-	-	-	-	2
FD	-	-	-	-	-	-	-	-	-	-	1
M	-	-	2	-	-	-	-	-	-	6	4
N	-	-	1	1	1	3	1	-	-	-	98
O&P	-	-	-	-	-	-	-	-	-	-	4
P	-	1	-	-	-	2	-	-	-	-	3
PR&R	-	-	-	-	2	2	-	-	-	-	8
RMBES	-	-	-	-	1	-	-	-	-	-	-
RA	-	-	-	-	-	-	-	-	-	-	1
RR	-	-	-	-	-	-	-	-	-	-	3
SD	-	-	-	-	-	-	-	-	-	-	-
SCR	-	-	13	-	-	-	1	-	-	-	6
TDIC	-	-	5	-	-	-	-	-	-	1	1
TG	-	1	-	-	-	-	-	-	-	-	-

Figure 5.1: Confusion matrix of labels over the four documents Dr. Bender and I annotated. Columns indicate labels annotated by Dr. Bender; rows indicate labels I annotated. Labels in agreement are highlighted in yellow. Instances of high label confusion ($n \geq 5$) are highlighted in red, and paired cells of these instances are highlighted in blue. Continued in Figure 5.2.

Hawaiian Language, 2019), and from a research institution (Ontario Federation of Indigenous Friendship Centres, 2016), selected randomly from those categories. I computed the inter-annotator agreement with Cohen's Kappa (un-weighted) (Cohen, 1960) using the scikit-learn python package (Pedregosa et al., 2011) (version 1.2.2). Across the four documents, the Kappa value is $k = 0.37$, where 0 indicates no agreement and 1 indicates complete agreement between two annotators across the labeled data. This relatively low inter-annotator agreement reflects the lack of contrast between the descriptions of the labels that stems from adapting a coding manual from the ISE Code of Ethics principles that were not intended to be coding manual labels. To look more closely at the agreement for each of the labels, the confusion matrix in Figure 5.1 and continued in Figure 5.2 show the counts for each label, with the rows showing labels produced by myself and the columns showing labels produced by Dr. Bender.

Label	O&P	P	PR&R	RMBES	RA	RR	SD	SCR	TDIC	TG
A&DC	-	-	-	-	-	-	1	-	-	1
AP	-	-	-	-	-	-	-	-	-	-
C&AP	-	1	-	1	-	3	-	1	-	-
C&II	-	-	-	-	-	-	-	-	-	-
C	-	-	-	-	-	-	-	-	-	-
D	-	5	-	-	-	-	-	1	-	-
D&R	-	-	-	-	-	3	1	1	-	-
FPIEC	5	-	-	-	-	-	1	-	-	-
FD	-	-	-	-	-	-	-	-	-	-
M	-	1	-	-	-	-	-	2	-	2
N	3	1	-	4	-	3	2	1	-	3
O&P	5	-	-	1	-	-	-	-	-	-
P	2	5	-	-	-	1	1	-	-	-
PR&R	6	-	-	-	-	-	-	-	-	-
RMBES	-	-	-	1	-	-	-	-	-	-
RA	-	-	-	-	-	-	-	-	-	-
RR	-	-	-	-	-	2	-	1	-	-
SD	-	-	-	-	-	2	4	-	-	1
SCR	-	-	-	5	-	7	7	10	4	3
TDIC	-	-	-	2	-	2	-	1	4	-
TG	-	-	-	1	-	1	3	-	-	14

Figure 5.2: Confusion matrix of labels over the four documents Dr. Bender and I annotated. Columns indicate labels annotated by Dr. Bender; rows indicate labels I annotated. Labels in agreement are highlighted in yellow. Instances of high label confusion ($n \geq 5$) are highlighted in red, and paired cells of these instances are highlighted in blue. Continued from Figure 5.1.

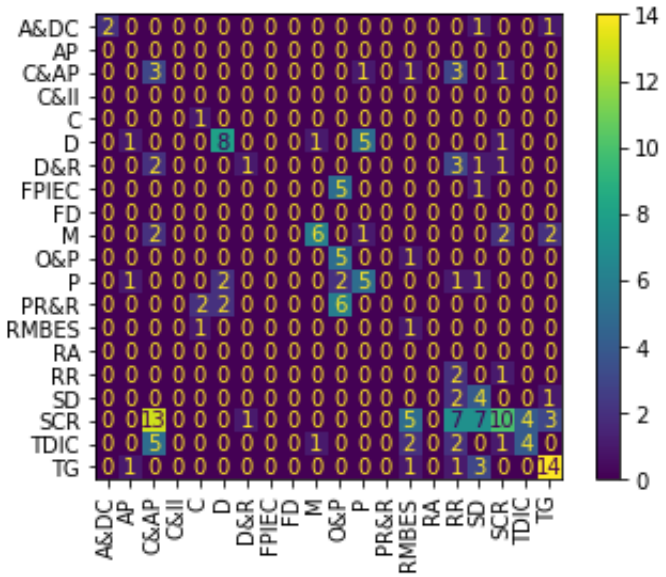

Figure 5.3: Heat map of the label agreement with *None* removed.

The *None* label is again the most frequent by a significant margin and has the highest agreement (98 instances), suggesting that it is clear when a sentence is not communicating one of the labels. Removing this category reduces the the Kappa value to $k = 0.31$. Figure 5.3 shows the instances of agreement and disagreement with the *None* label removed. The label with the highest agreement after *None* is *Traditional Guardianship* with 14 instances of agreement and no consistently interfering labels. *Traditional Guardianship* was the only label with a description that held the community knowledge as the primary focus, as opposed to the community members or outside researchers, so the labeling was relatively consistent between the annotations.

There are several interesting points of disagreement among the two sets of annotated labels. Of all the labeling pairs, *Diligence* and *Precaution* have the most confusion, in that there were multiple instances of both annotators labeling *Diligence* and the other labeling *Precaution*. There were 5 instances of my labeling *Diligence* and Dr. Bender labeling *Precaution* and 2 instances of Dr. Bender labeling *Diligence* and me labeling *Precaution*. The descriptions for both *Diligence* and *Precaution* focused on the actions that outside researchers could take to anticipate and prevent harms to the community prior to them occurring. While the description for *Diligence* highlights the harms resulting from an outside researcher's lack of familiarity with the community prior to engaging in research with the community, the description for *Precaution* highlights

the harms from the research itself. This distinction between the outside researcher and the research can be difficult to disentangle however, such as when an outside researcher's lack of familiarity leads them to frame research questions in such a way that disempowers community members. A label combining the two could emphasize this connection between the outside researcher's understanding of the community and the possible negative outcomes of the research that would need to be guarded against or mitigated. Collapsing *Diligence* and *Precaution* into one super-label raises the Kappa value to $k = 0.39$.

Similarly, *Prior Rights and Responsibilities* and *Ownership and Permission* have a high degree of confusion. While there were no instances of Dr. Bender using the *Prior Rights and Responsibilities* label, there were 6 instances of my using *Prior Rights and Responsibilities* and Dr. Bender using *Ownership and Permission*. The descriptions for these labels focus on the legal implications of community knowledge. I added the *Ownership and Permission* label to distinguish the permitted uses for third parties defined in the analyzed licenses from the rights that community members themselves may have in using their knowledge. However, several of the annotated instances discussed the copyright of the material, which could be understood to both protect the legal rights of the owner of the copyrighted material and regulate uses of the material by third parties. A combined label could enumerate the different kinds of legal rights and protections available for community knowledge. Collapsing *Prior Rights and Responsibilities* and *Ownership and Permission* into one super-label also increases the Kappa value to $k = 0.39$. Using both super-labels (*Diligence* + *Precaution* and *Prior Rights and Responsibilities* + *Ownership and Permission*) raises the Kappa value to $k = 0.41$. These labels are highly dependent on which 10% of the document set are labeled by multiple annotators, but the increases in the inter-annotator agreement from collapsing the labels into super-labels could inform future coding manual development.

Supporting Community Research was the most inconsistent label. Although there were 10 instances of both Dr. Bender and I agreeing on a sentence being labeled as *Supporting Community Research*, there were also 14 instances of confusion with *Collaboration and Active Participation*, 8 instances of confusion with *Respectful Relationships*, 7 instances of confusion with *Self-determination*, and 5 instances of confusion with *Reciprocity, Mutual Benefits, and Equitable Sharing*, among others. This overlap was in part due to my editing *Self-determination* to include determining research directions, but it is also because supporting community research implicitly requires all of the other principles, especially collaboration, respectful

relationships, community self-determination, and reciprocity. While merging *Supporting Community Research* with another label may have reduced the overlap, it is less clear which label it should be merged with. Instead, focusing on the 10 instances of agreement, a better path forward may be honing the *Supporting Community Research* label to capture what the other labels miss. Dr. Bender noted that the *Supporting Community Research* seemed appropriate even when extending to activities beyond research, and, looking at our agreed instances, the core of this label indeed focuses less on research and more on community capacity-building, sustainability, and transferring skills to the community in all contexts. Renaming and refining the *Supporting Community Research* to, e.g., *Community Capacity-Building and Sustainability* may produce more consistent inter-annotator agreement.

5.4 Lessons from the Annotation Analysis

The trends across the document types (§5.3.2) present general patterns of which labels are more frequently emphasized throughout the documents. As discussed in §5.3.1, the labels group together best practices towards normative goals in support of community collaboration in research. The labeled data then consists of support for and examples of similar best practices across the communities. In this section, I provide my own interpretations of the values these actions may be supporting, with many values finding support across several different sets of best practices represented by the coding labels. In particular, I focus on values necessary for building collaborative relationships with communities to inform the development of C3DAR as a tool for supporting collaborations.

Self-determination More than half of the analyzed documents recommended best practices that support self-determination. Self-determination is called out explicitly by **SL2**, **EU4**, **SA3**, and **OC2** and made reference to implicitly in other documents through emphasis on communities' decision-making processes. Community self-determination can happen through many different kinds of processes, such as within families (e.g., **SA3**), through elected boards and councils (e.g., **AF3**, **NA4**, and **EU4**), and by community leaders (e.g., **AS2**). In community research, support for self-determination is evidenced by research that will "produce direct benefits to Aboriginal people and reinforce Aboriginal peoples' self-determination through their full and ongoing active participation and negotiation in the decision-making process for research planning and im-

plementation according to local priorities," as suggested by **OC2**. As advocated for by **NA2**, taking a collaborative approach to research rather than a participatory approach also affirms community self-determination by meeting community researchers on equal footing and including community-defined goals within the research objectives. Consent for research activities can also support community self-determination when it goes beyond simply presenting research participants with a consent form. **AF3** states that consent is "also a process in itself, and one by which the Khoikhoi are able to conduct their own independent and collective discussions and decision-making." Within these processes, **OC5** suggests individual consent and collective consent may both need to be collected depending on the community protocols and possible risks of the research. Furthermore, the documents showed that self-determination is important for community maintenance of traditional knowledge and future data governance policies. Acknowledging and supporting self-determination is the only way to begin to repair academic and community relationships that have historically resulted in researchers extracting community knowledge and dispossessing communities of their rights to their epistemologies and their languages.

Recognition For both individual community members and communities as collectives, recognition is a relationship-affirming value. Recognition of individual work within a research project is generally required for ethical practices, but it is especially important for communities whose contributions to research have often been left out of publications. That being said, co-authorship is not always the most appropriate form of recognition for certain community members (e.g., in communities where an individual standing out from or speaking for the community may be considered an offense as suggested by **NA2)** and many guidelines recommend open discussions of alternative ways to recognize contributions to research (e.g., **SL2** and **NA4**). Within research dissemination, communities appreciate recognition of community variation and diversity, both for addressing the unique needs within the community (e.g., **AF4**'s language learning materials for diaspora groups) and in order to not privilege one group over another by singling them out as *the* community representative (e.g., **AS2**'s description of the recorded stories as the narrator's version, not the only or correct version). Communities' collective recognition of their individual community members belonging to the community is central to communities maintaining their governance structures (e.g., **AF5**). External recognition of communities, their right to self-determination, and their self-determined needs contribute to more research collaborations using community ways of knowing, more visibility and resources for that re-

search, and legislation that may support the community in realizing their goals. For example, **AF1** describes signed language linguists as "important allies for Deaf associations in their campaigning for equal rights for deaf people, including legal recognition for national sign languages and their facilities (interpreters, access to education, information, etc.)" and **SL3** argues for "visibility and recognition of sign languages as natural human languages" as a primary goal for research with signed language communities. Importantly, external recognition in and of itself is not enough to empower communities, however it is a first step in building relationships with communities and community members grounded in respect.

Respect Respect is closely tied to other values like empathy and care. **AF5** frames respect as the acknowledgement of community research contributions at all times, engagement with communities in advance of carrying out research, protection for communities' and community members' privacy, and care for the families of those involved, as well as to the social and physical environment, in addition to the direct research stakeholders. While academia has discouraged researchers forming relationships with communities they work with on the basis that research in such conditions lack objectivity (Haualand, 2017), **AF5**'s suggestions show that having respect and compassion for community members is often orthogonal to the research question itself. **AF5** and other communities challenge the claims that objective research is inherently better; **AF5** contends that research "lacking in care for the community" is "not up to a high standard" and may result in negative interactions with the community. For **OC5**, "cultural and social responsibility and respect for persons" and acknowledgement of "a person's inherent dignity and the responsibility that people have to act in a caring manner towards others" are simply a minimum standard for ethical research. Going beyond minimum standards, **OC5**, **NA2**, and **NA3** argue that engaging in research methods and community interactions based in empathy and active efforts to understand communities' experiences result in more effective research grounded in communities' own perspectives, rather than external perspectives that may misunderstand their situations. **SL2** also supports respect for "the diverse experiences, understandings, and way of life (in sign language societies) that reflect their contemporary cultures"; such respect contributes to respect for communities' self-determination by not imposing normative expectations of how the communities "should" use their language or live their lives. Community researchers are known by their communities and generally have respect and empathy for their community; established relationships help to support community research collaborations like the project described in **AF2**, who report that their community "fully support and trust

them and are happy to work with them." Respect and empathy are relevant to so many aspects of research because they are foundational to ethical research regardless of the research topic, methods, or results.

Honesty Like recognition and respect, honesty is necessary for growing and maintaining strong relationships with communities. Honesty itself is based in ongoing dialogue and communication that builds understanding between the research team and the community by using language(s) (e.g., **OC5**), media (e.g., **NA4**), and terminology (e.g., **AF5**) that are accessible to the community. According to **OC5**, for Māori communities "a relationship displaying transparency, good faith, fairness and truthfulness is captured in the concept of whakapono (hope) and the whakatauki (proverb) 'kia u ki te whakapono, kia aroha tetahi ki tetahi' (Hold strong to your beliefs and care for one another)." **AF5** also note the relationship between honesty and transparency, stating that, "Honesty also means absolute transparency in all aspects of the engagement, including the funding situation, the purpose of the research, and any changes that might occur during the process." Generally, communities want to engage in dialogue towards the mutual exchange of information (e.g., **NA2**), but the responsibility to be transparent and fully disclose information lies only with the research team. In this way honesty and transparency differ: honesty is about building shared understanding between people by sharing information believed to be true by the sharing party, while transparency is about providing as much truthful information as possible. Honesty and transparency therefore have an asymmetric relationship where being transparent also entails being honest but being honest does not entail being transparent. In order to protect sensitive community knowledge and vulnerable community members, communities may opt to not share community information as part of the research and this should not be seen as contradicting honesty. Instead, honesty in these cases (such as **SA3**) is simply stating that the information cannot be shared.

Honesty is often necessary for protective measures like consent and precautions related to the potential impacts of research. For **AF5**, free, prior, informed, and educated consent "can only be based on honesty in the communications, which needs to be carefully documented." **OC5** lists many considerations for what needs to be conveyed in the consent procedure, including "the potential or real risks" involved in the research. These considerations are important for both protecting individuals and protecting the community. Failure to be honest when communicating with communities about the scope of potential harms or even the range of possible benefits to the community can be seen as a "violation of community trust" according to **AF5**.

OC5 emphasizes the "importance of the researcher's own credibility, trust, honesty and integrity vis-à-vis the research project and participants" in maintaining ethical collaborations with communities in addition to increasing infrastructure for ethical research like community guidelines and ethical review processes.

Reciprocity Recommended actions to support reciprocity were especially highlighted across the documents. Motivations for reciprocity included fairness (e.g., **NA4**), respecting community self-determination and supporting community goals (e.g., **OC2**), and restitution after years of exploitative research practices (e.g., **AF3**). While researchers benefit from research with communities in terms of academic achievements like status and reputation, qualifications, personal career advancement, and increasing networks, often more intentional effort is needed to ensure communities also benefit from the time, knowledge, and resources they provide to the research, according to **OC5**. This has led to the formation of community biocultural protocols like the ones described by **AF3** and **SA3**, which protect community traditional knowledge and manage economic benefits from that knowledge such that the benefits are equitably distributed to the community members. For **SA3**, "the practices of voluntad (willingness), ayni (mutual assistance) and minka (exchange of labour)" define reciprocity. In the context of research, willingness aligns with enthusiastic consent to be part of the collaboration (e.g., **AS2**). Mutual assistance might look like all collaborators sharing their knowledge and their skills (e.g., **SL2**). Exchange of labor within research could manifest as all of the collaborators actively participating in the design, data collection, and analysis phases of the research (e.g., **SA4**) or could manifest as community members contributing to research in exchange for the researcher supporting community goals, either through the current research project or in some other way. For example, **AF4** describes how a language documentation project to produce a grammar and standardized orthography for the Sakun community provides the foundation for future work toward developing pedagogical materials, including second-language learning materials for community members living away from the community. Reciprocity contributes to the relationship with the community by evincing the researchers' care for the community and for the community's self-identified needs.

Responsibility Community researchers, as community members, have a strong sense of responsibility for how their research impacts the entire community. By conducting research with communities, the guidelines and licenses argue that outside researchers take on their own responsibilities towards the community. When

initiating a collaboration with a community themselves, **NA4** and **SL1** recommend that outside researchers first learn about the community's customs, history, and language. **SL2** states that proficiency in the language of the community is especially important as "use of interpreters negatively impacts the project" and hinders communication. Then, **NA4** and **SL1** recommend that researchers investigate current and prior relevant research with the community in order to not reuse community time and resources for repeating prior work. After this preparation, informal introductions with community members can lead to new and mutually beneficial projects which can then be presented to ethical review boards and community authorities for approval. Some communities like **EU4** require an application process showing that the research team has enough knowledge on "Sámi health, traditions, history, Indigenous knowledge, and social situation" prior to approving research projects with the community. Once the research starts, **OC5** then advises that, "The responsibility to protect and care for people with aroha and be aware of issues of cultural sensitivity comes to the fore," including ethical deliberation of possible research outcomes and considering "the nature of the outcomes (risk versus benefit, short versus long term) and their relative distribution (researchers, participants, communities, society)."[13] Honestly communicating the risks and taking all precautions to mitigate them is another responsibility that falls to researchers, both inside and outside the community. Once the research has been safely completed, **OC2** states that researchers have the responsibility to return research outputs, data, and community-appropriate disseminations of the research results to the community. Even for researchers who are just using the community data after it has been collected, there is still the responsibility to adhere to the terms of use of data licenses, protect the privacy of those in the dataset, make sure that attributions properly credit the community's contributions, and return the results of the research to the community. Fulfillment of these responsibilities builds trust within the collaboration and shows that the community can rely on the research team to care for and respect the community members.

Privacy Privacy applies to both individuals and to communities. Individuals may need privacy for their personal information like biometric data, health data, and even their name. In addition to protecting individuals from identification and subsequent negative impacts, **AF5** argues protecting individual privacy is grounded in respect for individuals. This respect may be for the inherent dignity of each person, and it may be for the ability of that person to determine for themselves how their data may or may not be used.

[13] **OC5** glosses the Māori term aroha as both *care* and *awareness* but note that neither fully captures its meaning.

Anonymization of the data can help protect individuals, but needs to be done carefully to ensure that the data is still useful to the community. Signed language communities especially stress the care that must go into conveying the risks of video data to community participants when asking for consent and then anonymizing signed language video data (e.g., **SL2** and **SL3**). Learning about the linguistic factors that contribute to challenges with anonymizing signed language data and protecting signers' privacy is one of the responsibilities of signed language researchers that contribute to a sense of care in the collaboration. Community data may require privacy measures for traditional knowledge that needs to be protected or for sensitive community topics, though how to handle these kinds of data is best decided in consultation with community leaders according to **OC5**.

Privacy is also a primary concern of data archives. All of the archives I examined provided at least one level of data restriction (openly accessible or restricted) for individual files within a data collection submitted to the archive. ELAR for example provides three levels of restriction: 1) openly accessible, 2) access to archive subscribers who have agreed to the archive code of conduct, and 3) access only to individuals who have requested permission from the data manager. Archives vary in terms of remedial actions they will take following user violations of the archive codes of conduct, such as breaching community privacy. AILLA will notify the AILLA community of the violation, while Kaipuleohone and PARADISEC have terms that authorize them to make decisions on behalf of communities in such cases. These different courses of action show a tension between respecting community self-determination and taking active measures to protect the community. While AILLA's course of action allows communities to determine their own responses to privacy breaches, communities may not have the resources to hold users accountable. Kaipuleohone and PARADISEC's course of action, on the other hand, prioritizes holding users accountable at the expense of possibly including communities in the accountability process.

Accountability Accounting for one's actions can happen in two different ways. On the one hand, **SL1** and **OC5** suggest researchers can be accountable to communities and other overseeing institutions through honest and frequent communication about the project, documentation of the project, and upholding community expectations with respect to researcher responsibilities. Willingness to share information and be accountable to communities supports trust and transparency within the collaboration. In addition to the community, researchers may also be accountable their affiliated institution and any funding agencies. **NA2** warns that

the expectations and turn-around times of these differing accountability structures may pose challenges for collaborative teams and to be aware of how they may impact each other and the research timeline.

On the other hand, researchers can be held accountable for their research and its impacts on the community. Collaborators could potentially hold each other accountable through reflective processes (e.g., **NA3**) or by refusing to continue working with other collaborators until they fulfill their responsibilities (e.g., **AF5**). Communities also have methods for holding researchers accountable. In some instances, **AF5** states that institutions whose researchers fail to comply with the community code of ethics may be refused future collaboration opportunities. Community oversight committees and review boards can also enforce accountability measures where they exist (e.g., **OC5**), however not all communities have the power and resources to establish their own accountability infrastructure. For both individual collaborators and communities, asymmetrical power structures may hinder holding researchers with power accountable for their impacts on communities and on researchers with less power within the collaboration.

Awareness Being aware of one's position with respect to the community and with respect to the other collaborators can have a significant impact on the research and on the collaboration itself. Positionality can be literal; **NA4** reminds researchers to consider that, "wherever you are conducting your research, you are in someone's community or a First Nations territory." **OC5** also conceptualizes awareness of one's position as "acknowledging the essence of the environment within which a person operates." Both convey the sense of no longer being in one's own space and needing to treat the space you've been welcomed into with care. Awareness may also mean acknowledging different perspectives and expectations that come from lived experience within the community or in close proximity to the community. **EU4** points out that, "Sámi researchers doing research in their communities already have a standing as members of their communities, and thus the issues they need to consider are different in comparison to researchers coming from outside the communities." Similarly **SL1** also suggests that "peer-support from a neighbouring Deaf Community might be more fruitful than from deaf partners originating from a very different socio-cultural and politico-economical context" in research projects. Awareness is thus important for outside researchers to practice in order to understand where differences in assumptions and worldviews may be contributing to a different working relationship with community members or even a different research interpretation, as suggested by **NA2**.

Awareness can also contribute to assessing power dynamics in the research collaboration and changing those dynamics if necessary. **NA3** and **SL2** suggest reflective processes and critical examinations for doing so. Researchers have typically been in positions of power relative to indigenous communities (as noted in e.g., **NA3**) and signed language communities (as noted in e.g., **SL3**). Even when unintentional, power dynamics will often settle in these arrangements due to surrounding systems built on oppression; being aware of power dynamics and actively working against them is the only way to change power dynamics within the collaboration and within general research practices.

Justice Justice is about treating people equitably and repairing the harms of systematic oppression over generations. For **AF3** and the Khoikhoi people, the harms include "the last 150-200 years" of exploitation, intellectual theft, and violations of consent: "we as the original knowledge holders never benefited from sharing our knowledge, we never received intellectual property rights, nor was our free, prior and informed consent given." Righting those wrongs requires not only the cessation of exploitative practices by external governments and industries, but also lifting communities back up from oppression through "identifying tangible outcomes for all parties and supporting more equitable benefit sharing" (**OC5**) and community capacity-building through training and collaboration (e.g., **AF5** and **SL1**). Justice and community self-determination are interdependent; these actions towards justice for communities also support their ability to determine their own futures, which further helps in identifying community needs and quickly mitigating new instances of harm. **NA3** asserts that only through recognition and acknowledgement that indigenous people "have been, and remain, disfranchised, disadvantaged, and dispossessed" can indigenous communities regain full control over "generation and dissemination of knowledge" and "make decisions about their lives, assert their rights to execute plans, goals and priorities, and own their cultural, socio-economic, and political reality." This control over generation and dissemination of knowledge supports communities' abilities to identify possible harms in new research proposals through community-established review boards and mitigate those risks before they impact the community, as exemplified by **OC5**. These processes support communities' ability to hold researchers accountable for the impacts their research has on the community in ways that were not possible when communities' rights were not recognized. Community control over data and knowledge also helps to prevent harms from systems of oppression in a new digital landscape; **OC4**, in its rationale for its data license, argues that, "By simply open sourcing our data and knowledge, we fur-

ther allow ourselves to be colonised digitally in the modern world." By developing infrastructures and tools to support community self-determination, communities are able to work towards more just and equitable futures.

Summary In this section, I discussed ten values central to collaboration and relationships that were interwoven in the text of the documents: self-determination, recognition, respect, honesty, reciprocity, responsibility, privacy, accountability, awareness, and justice. Each of these values connected several of the labels representing broad categories of recommended best practices and found support in recommendations from different regions and different language communities. While guidelines tended to address values more directly, licenses and archives documents also provided considerations for collaborations related to responsibility and privacy in digital spaces.

5.5 Limitations and Next Steps

In this chapter, I have presented a retrospective technical investigation analyzing community documents for ethical research and engagement with communities. To learn from the stated values of language communities with respect to their data, I gathered and selected diverse guidelines and licenses from prior work and archives for language data. The collected documents described in §5.1 represent perspectives from communities and research organizations around the world. I selected a final set of documents (§5.2) such that each geographic region included the considerations of signed language communities, a variety of community documents, and both licenses and guidelines.

Although the documents were selected to be broadly representative, this work and the prior work it builds on is restricted to documents that have been made publicly available and addressed to third parties outside of the community. As Hayward et al. (2021, pg. 411) point out, "informal, oral, and/or non-publicly facing Indigenous research ethical protocols and frameworks" could provide additional insight into community perspectives on ethical research. While Tunón et al. (2016) looked at guidelines published in Swedish, the vast majority of the documents examined in my work and the work it builds on is in English. Future work in community research ethics may provide more insight by considering protocols developed for the authoring community itself and in the authoring community's language.

Using a coding manual drawn from the ISE Code of Ethics text (International Society of Ethnobiology, 2006 with 2008 additions), I annotated each sentence of the documents selected for the values they express as defined by the labels in the coding manual. My results (shown in §5.3) point to *Supporting Community Research*, *Diligence*, and *Reciprocity, Mutual Benefit, Equitable Sharing* as frequently mentioned values across all of the documents. I also compared my annotations on 10% of the documents to a second annotator's in §5.3.3. This comparison indicated that my changes to the coding manual were conservative. The relatively low inter-annotator agreement reinforces the benefits of having an interdisciplinary team contribute to developing the coding manual. The discussions that my co-authors and I had in refining our coding manual in Chapter 3 for coding the workshop data led to a more honed set of labels. Despite these drawbacks, I leveraged the points of disagreement to suggest further modifications of the coding manual to support future annotation efforts. Close analysis of the guidelines and licences during the annotation process also yielded general patterns across the guidelines and licenses regarding best practices for research conducted with communities.

From these patterns, I discuss key values supported by the recommended best practices and approaches to ethical community engagement as expressed by communities themselves in §5.4. The breadth of lessons from the documents reveal interconnected values focused on building relationships that are carried through the research design, methods, publication, and after the research is completed. In Chapter 6, I conduct a technical investigation to integrate these lessons into data statements, creating the first iteration of a generalized toolkit for collaboratively developing community language datasets.

Chapter 6

Technical Investigation: Developing C3DAR

In this technical investigation, I draw from three sources to develop the first version of C3DAR. The first is data statements Version 2 (DSV2) as presented in Chapter 3. This provides the foundation from which I build C3DAR. In §6.1 I begin by shifting DSV2 from a retrospective documentation toolkit intended for documenting datasets that have already been collected to a dataset design and planning perspective in which the data has yet to be curated. Following these changes, I make use of the parts of the comparison with datasheets from Chapter 3 that were deemed out of scope to add important considerations for data longevity, namely maintenance and distribution, to the schema (§6.2). Finally in §6.3 I apply the lessons learned from the retrospective investigation in Chapter 5 to the working version of C3DAR, resulting in the first version of C3DAR. The full text for the C3DAR toolkit is presented in Appendix B and is available for download along with a template at `https://digital.lib.washington.edu/researchworks/handle/1773/50585`.

6.1 Repurposing Data Statements

Throughout the data statement schema, the directives for each element assume that the dataset developers have already completed the collection and annotation of the dataset and are now in the process of publishing the dataset along with appropriate and thorough documentation. The intended use case for C3DAR is as soon as a collaboration has been agreed upon and before any data collection occurs. The C3DAR schema can be used to detail the dataset design and intended goals at the start of the project, then the documentation

should be updated to reflect any changes to the actual dataset development. This new purpose, in part inspired by the comments from the participants of the workshop described in §3.5.1, was also suggested as a possibility in DSV2 general best practice (GBP) 3: "Consider using the data statement elements as a checklist for dataset design."

As a GBP, this statement becomes redundant, as the consideration is now part of the main purpose of C3DAR, in addition to supporting collaboration. The DSV2 template presents its purpose with an introductory descriptive statement:

> A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software. In developing and writing your data statement, keep in mind that someone new to your dataset might be reading your data statement 10 or 20 years from now. Alternatively, at some point in the future, you might find yourself looking back at your data and reading your data statement to recall key characteristics and decisions.

I remove this statement and the GBP on using the elements for dataset design and instead begin C3DAR with the following statement:

> This toolkit is designed to support collaborative dataset curation and documentation between language communities and technical communities. It consists of general best practices, a list of key terms, and 17 schema elements corresponding to key considerations for designing datasets and writing documentation. Each schema element includes the rationale for its inclusion in the schema, its definition, and suggested best practices. By filling out each of the schema elements with a future dataset in mind, the dataset design team can thoroughly discuss plans for the dataset's content, creation process, and publication while also producing an initial draft of the dataset documentation. This process is intended to be iterative, with schema elements being drafted as decisions are made and updated as the project develops.

This statement emphasizes C3DAR's primary uses in collaboration facilitation and dataset design while also framing documentation as an additional product.

I took an additive approach to the remaining GBPs, the key terms, and the schema elements, in that I did not exclude any of the original sections. However, they still required some modification to fit with the new intended purpose. First, I removed named instances of data statements throughout the schema, changing the wording to instead reference "documentation." In some instances, "data statements" was self-referential and used to indicate the instance of a data statement at hand, as in 1 Header. In this element, "Data Statement Author(s)," "Data Statement Version," and "Data Statement Citation" all referred to the author(s), version, and citation of the dataset at hand (rather than, for example, the authors of the data statement template). These were changed to "Documentation Author(s)," "Documentation Version," and "Documentation Citation." In one instance this change from "data statement" to "documentation" made a schema element description non-sensical. The *What* section of 4 Documentation for Source Datasets stated, "For datasets built out of pre-existing datasets, a link to a data statement for each source dataset should be included. If a data statement is not available, provide a link to a publication or other documentation." Simply replacing "a data statement" to "documentation" would have caused confusion as to what other documentation should be linked to in the case that documentation in general is not available. Instead, I changed the description to, "For datasets that will be built out of pre-existing datasets, a link to the documentation for each source dataset should be included." This phrasing makes it clear that any available documentation for source datasets will suffice.

The third step for repurposing DSV2 required changing the perspective of the documentation with respect to the documented dataset. In DSV2, the dataset has already been created and therefore is referred to in the past tense. In the intended use case for C3DAR, the dataset has not yet been created and therefore is referred to in the future tense. This largely involved simple instances of changing verbs to future tense such as GBP 2: "For dataset containing sensitive or proprietary information..." becomes "For datasets *that will contain* sensitive or proprietary information...". The *What* section of 3 Curation Rationale required an additional edit. The first question asked, "Why was this dataset created?" edited to the future tense, "Why will this dataset be created?" could be interpreted as trying to request a broader rationale for the collaboration itself. To clarify that the focus is strictly on the dataset, I changed this question instead to "What is the intended purpose of this dataset?".

These structural modifications to the perspective of DSV2 formed the basis of the first version of

C3DAR. To add more to the content of C3DAR, I turned to another documentation format as we had for DSV2, datasheets for datasets (Gebru et al., 2021). The sections of datasheets that we had determined were out of scope for DSV2 provide important future-looking considerations for the design of datasets.

6.2 Incorporating Out of Scope Topics from Datasheets

In §3.5.2, my co-authors and I compared our intermediary version of data statements with datasheets for datasets. From this comparison we identified topics that were addressed in both data statements Version 1 and datasheets as well as a number of topics that were addressed in datasheets but were not considered in data statements. The comparison with datasheets for datasets (Phase 2) yielded five additions; all of these were to element descriptions. We left several considerations from datasheets as out of scope for the intended use case of DSV2; these were mostly from the sections titled *Uses*, *Distribution*, and *Maintenance*. Whereas we used v7 of datasheets in the comparison with DSV2 (the latest available version at the time), I used the version of datasheets as published in the *Communications of the ACM* for this investigation. The text from the original sections in datasheets and their integration into C3DAR are presented in the following sections.

6.2.1 Original Questions from Datasheets for Datasets

As discussed in §3.5.2, datasheets were the most appropriate available documentation format for comparison with data statements. Both target datasets, as opposed to data and models, for documenting, and both use a prose format for presenting information rather than a visual format. Differences in the development contexts of datasheets and data statements (industry and academy, respectively), conceptualizations of data, and the intended audience suggested opportunities for different dataset considerations to appear in each schema. Comparing the content of datasheets and DSV2 allowed us to determine if there were any significant dataset details that we had not yet considered including in the DSV2 schema. We found that each of the questions in the datasheets sections had a relevant equivalent question in data statements with the exceptions of the *Uses*, *Distribution*, and *Maintenance* sections.

Uses contains a rationale and six questions about how the dataset is being applied and how it might be applied by third parties:

The following questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- 37. Has the dataset been used for any tasks already? If so, please provide a description.
- 38. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
- 39. What (other) tasks could the dataset be used for?
- 40. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
- 41. Are there tasks for which the dataset should not be used? If so, please provide a description.
- 42. Any other comments? (Gebru et al., 2021, pg. 91)

Distribution consists of a brief instructional statement and seven questions about the entities and methods involved in distributing the datasets:

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

- 43. Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

- 44. How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
- 45. When will the dataset be distributed?
- 46. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
- 47. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
- 48. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
- 49. Any other comments? (Gebru et al., 2021, pg. 91)

Maintenance contains both a rationale and instructions as well as eight questions about the individuals involved in managing the dataset and communication of dataset updates and changes:

> As with the previous questions, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

- 50. Who will be supporting/hosting/maintaining the dataset?
- 51. How can the owner/curator/manager of the dataset be contacted (for example, email address)?
- 52. Is there an erratum? If so, please provide a link or other access point.
- 53. Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

- 54. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
- 55. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
- 56. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
- 57. Any other comments? (Gebru et al., 2021, pg. 91)

Throughout these sections, we noted some questions had partial overlap with the content in 1 Header, 3 Curation Rationale, 11 Limitations, and 12 Metadata. For *Uses*, we mapped Question 37 to the content of 3 Curation Rationale and Question 41 to the content in 11 Limitations. For *Maintenance*, we mapped Questions 44 and 45 to 1 Header and added a prompt for a DOI in DSV2 as a result. We also mapped Question 46 as well as Questions 54 and 56 in *Distribution* to our questions in 12 Metadata. We added a suggestion for linking to a list of errors found after the dataset's publication (Question 52) to 12 Metadata as a result of the comparison. The remaining questions were deemed out of scope for DSV2. For C3DAR, however, the remaining questions touched on important topics for communities and researchers to discuss with regards to the ongoing care and resources that will go into the dataset as evidenced by the frequent mentions of ownership, permissions, access restrictions, and maintaining community control over data in Chapter 5.

6.2.2 Incorporation into C3DAR

In deliberating how to incorporate the content from the datasheets sections into C3DAR, I weighed the decision to create new schema elements against how much content remained to be incorporated after the

changes we made in §3.5.2. Question 38 is about past uses of the dataset by other parties, and so that question is out of scope for C3DAR as it cannot be answered at the time when communities are still designing the dataset. The remaining questions about intended uses are covered by 3 Curation Rationale and potential impacts on later uses as a result of design decisions have a space in 11 Limitations. However, the main questions for *Distribution* (mainly, will the dataset be distributed and if so where and under what terms) and *Maintenance* (who is going to be responsible for the dataset) are not yet represented in C3DAR. For these reasons and mindful of the length of C3DAR, I incorporated only *Distribution* and *Maintenance* as new schema elements.

To adapt the datasheets questions to the DSV2/C3DAR style, each new schema element needed a *Why* section providing a rationale for documentation authors and readers, a *What* section describing what content to include (rather than asking questions), and best practices. I started by translating the questions into a schema element *What* description. Then I drafted the *Why* sections using the rationales available in datasheets and with inspiration from the guidelines and licenses examined in Chapter 5. I left the best practices to be developed from Chapter 5's retrospective investigation lessons, discussed further in §6.3.

Development of the *What* Sections

The main content of the schema elements is presented in the *What* section. Datasheets were developed with technical professionals in industry and government in mind as the intended users. For C3DAR, the content of the questions also needs to be accessible to communities, academics, and professionals who might not have technical knowledge related to NLP dataset creation. The text from the datasheets questions therefore needed to be edited for style, format, and technical terminology.

Distribution The first sentence begins with a statement of the purpose of the schema element, which is describing the how the dataset will be distributed. I rearranged the remaining content to follow from this purpose. This meant bringing to the forefront the methods for distributing and/or restricting the dataset. In addition to the methods presented in datasheets, I added distribution through a data archive and access restrictions on a subset of a dataset. Furthermore, I added a suggestion that these restrictions may be on the basis of the data being deemed sensitive or confidential. The terms of use and other legal mechanisms are the final recommended details to include in the description. While these legal mechanisms require a

specific kind of technical knowledge, the guidelines analyzed in Chapter 5 stress the importance of taking the time, *prior* to data collection, to make sure that both communities and their research partners understand who will own the data in all of the relevant senses described in Bragg et al. 2021, how the data may be used, and how those uses will be enforced. Further details that may be discussed but are not necessarily required include parties the dataset will be distributed to, conditions for accessing restricted portions of the dataset, the dataset digital object identifier, and when the dataset will be distributed. These are included as bulleted suggestions rather than in the main description of the element. The final *What* section for the new schema element on distribution considerations is as follows:

> A description of how the dataset will be distributed should be specified. This includes the method of distribution (e.g., through a data archive, files on website, API, GitHub) and any access restrictions on the dataset or subsets of the dataset (e.g. sensitive or confidential content, intellectual property (IP)-based restrictions, export controls, or other regulatory restrictions). If the dataset or portions of the dataset will be distributed under an IP license, copyright, or terms of use (ToU), describe the licenses, copyright, and/or ToU. Provide links or other access points to, or otherwise reproduce, any relevant licensing terms or ToU, and list any fees associated with these restrictions. Other suggestions for detailing the distribution plan include:
>
> - Who the dataset will be distributed to (e.g. third parties outside of the entity (community, company, institution, or organization) on behalf of which the dataset was created)
> - If there are conditions for accessing the dataset or subsets of the dataset, and if so, what the conditions for being granted access are
> - If the dataset will have a digital object identifier (DOI)
> - When the dataset will be distributed

Maintenance The *What* section of the schema element for maintenance considerations also needed an opening statement of what should be included in the element. While I made fewer changes to the content of the maintenance questions than the distribution questions, I only included how the dataset will be maintained, who will maintain it, and how to contact them as the main specifications for completing the element. The remaining questions were listed as additional considerations. Discussing who will be responsible for the

published dataset, *prior* to the data being collected, is important, however requiring further details that may not be knowable at the design stage may hinder the collaboration by suggesting that they should be known. The final *What* section for the new schema element on maintenance considerations is presented below:

> A description of how the dataset will be maintained should be specified. This includes who will support, host, and maintain the dataset and what the proposed method for contacting the manager of the dataset will be. Other considerations include:
>
> - If and where a list of errors found after the dataset's publication will be maintained and how to report errors
> - How often, by whom, and how updates to the dataset (e.g., to correct labeling errors, add new data, delete data) will be communicated to users (e.g., mailing list, GitHub)
> - Applicable limits on the retention of the data associated with the instances (e.g., will individuals in question be told that their data will be retained for a fixed period of time and then deleted) and how those limits will be enforced
> - Whether older versions of the dataset will continue to be supported, hosted, and maintained
> - How users will be notified that the dataset is outdated or no longer available
> - Whether others will be able to extend/augment/build on/contribute to the dataset, and if so, how others will be able to contribute, if and how these contributions will be validated, and whether these contributions will be further communicated and distributed to other users

Development of the *Why* Sections

The *Why* section of a schema element explains the rationale for why the schema element may be useful for the dataset creators or documentation authors as well as for the documentation readers. Datasheets do not systematically include these rationales with their questions, so I wrote rationales for both of the new schema elements based on reasoning from the DSV2 schema elements and from the analyzed guidelines and licenses in Chapter 5.

Distribution For other schema elements, such as 3 Curation Rationale and 6 Speaker Demographic, we cited intentional data collection practices as rationales. Knowing the distribution plan prior to collection may inform decisions against collecting sensitive data, for example as in cases where the dataset would be made publicly available or the development team could not guarantee the anonymity of the data subjects. Several of the guidelines (e.g., The Khoikhoi Peoples' Rooibos Biocultural Community Protocol (National Khoisan Council & Cederberg Belt Indigenous Farmers Representatives, 2019) and Te Ara Tika Guidelines for Māori Research Ethics (Hudson et al., 2010)) also stated that communicating dataset distribution plans to people whose data will be in the dataset was important to include in a free prior informed consent process. The dataset team must therefore have a distribution plan prior to beginning the collection process. The rationale for data creators includes both of these reasons:

> For dataset creators, having a detailed plan for distribution can help inform data curation decisions as it determines whether the team should only collect data that will allow for public distribution or if the final dataset will be distributed in a limited manner. The data collection team will also need to provide distribution information to the people the data is collected from to ensure that the data providers can provide their prior informed consent on how their data will be distributed.

Suitability to intended use cases is a frequently cited rationale across the schema elements. Knowing the terms of use that a dataset is distributed under is therefore very important for documentation readers, especially if there are disallowed uses or restrictions on redistribution. In 12 Metadata, we used the metaphor that documentation is the "front door" of the dataset, in that it may be how people interested in using the dataset are first acquainted with it. With this metaphor in mind, more easily accessible documentation also may provide answers to people who would like to use the dataset but are encountering access restrictions to the data without explanations as to why the data is restricted. The rationale for documentation readers consists of these reasons:

> For documentation readers, a detailed description of the permitted uses of this dataset can help in determining whether the dataset is suitable for a particular use case and whether the dataset can be further redistributed. If the documentation is the first access point for a reader, the

> distribution explanation can help the reader find and access the dataset or explain why they are unable to find or access the dataset. Documentation that communicates planned revisions or removal of the dataset in advance may also help documentation readers prepare for changes to the dataset.

Maintenance The description provided in Gebru et al. (2021)'s *Maintenance* section provides some rationale for dataset creators: "The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers." I adapted this rationale to also specifically name the community as an audience for whom a maintenance plan is important:

> For dataset creators, a maintenance plan for the dataset may help to ensure that the dataset will continue to be usable by and accessible to the intended audiences as well as the community that the data was collected from.

Again, informing use cases is a primary rationale for documentation readers. In the context of dataset maintenance, a documentation reader may want to ask questions of the maintainer or may be interested in whether different versions of the dataset exist. While the development team may not know this information at the time of the dataset design, they can update the documentation at a later date once the maintenance plan is finalized and the dataset is published. The rationale therefore serves as a reminder to review the documentation rather than a requirement at the time of the dataset development:

> For documentation readers, information about the dataset's maintenance will help determine who to contact for questions about the dataset after it has been published. Information about previous updates may help determine which version of the dataset will be most applicable to the reader's use case and help the reader plan for integrating dataset updates into their system development.

6.2.3 Reducing Overlap in Schema Elements

With the *Why* and *What* sections of the new schema elements for distribution and maintenance considerations written, the next step was integrating the new elements with the existing elements. Ethical review may

lead into discussions of distribution and maintenance if the review process requires a description of these plans. Discussions of funding will also have important implications for distribution and maintenance as these processes can necessitate significant human and infrastructure resources. For these reasons, I placed the new schema elements after 13 Disclosures and Ethical Review, making Distribution the 14th schema element, Maintenance the 15th, Other the 16th, and Glossary the 17th. These updated numbers were propagated to all references to those schema elements throughout the schema, such as the reference to Other in the general best practices.

Recall that content from the datasheets *Maintenance* and *Distribution* sections had been incorporated into the DSV2 schema. In order to reduce overlap between the schema elements, I consolidated relevant content in the new schema elements and removed it from the other schema elements. For example, 12 Metadata previously suggested including a link to the license and copyright permissions and a link to the list of errors found after the publication of the dataset. These suggestions were moved to 14 Distribution and 15 Maintenance, respectively. With the license information removed from the *What* section of 12 Metadata, the second best practice for 12 Metadata, to include license information for both the dataset and the data used to create the dataset, no longer fit. This best practice was removed, and suggestions to include license and copyright information for source datasets was moved to the *What* section of 4 Documentation for Source Datasets. Access restrictions to the data were included as part of 13 Disclosures and Ethical Review in DSV2; this content was removed from 13 Disclosures and Ethical Review in C3DAR since it is covered in 14 Distribution. Finally, I added a forward pointer to 15 Maintenance in the third best practice of 1 Header: "Consider web accessibility and the longevity of documentation location (e.g., university archives or a community-owned repository). See 15 Maintenance for further considerations." With these changes finalized, I turned to the lessons from the retrospective investigation to incorporate further changes to support communities, especially best practices.

6.3 Incorporating Lessons from the Retrospective Investigation

Further shifts in the phrasing of C3DAR were inspired by the lessons from the retrospective technical investigation in Chapter 5. The first changes were to the linguistics terminology used throughout C3DAR which defaults to spoken languages as the norm. Then I made changes to the schema elements where advice from

the ethical guidelines analyzed in the retrospective investigation could be applied. Not all of the advice was applicable to dataset documentation. Instead, it was aimed at the interactions between collaborators during the research. No schema element was appropriate for encapsulating these best practices, so I created a new set of general best practices for collaboration, presented at the beginning of C3DAR.

6.3.1 Bringing Signed Languages to the Forefront

In order to make C3DAR more welcoming to signed language communities, I made changes to the terminology used throughout the schema and added signed language examples in addition to spoken language examples when used in the element rationales. I started with the key terms, specifically *speaker* and *speech*. In linguistics, these terms are used to generally refer to a language user and instances of language in any modality, spoken, signed, or written. Using speech as the general term contributes to the lack of visibility for signed languages, so I removed uses of *speaker* and *speech* as general references to language regardless of modality and removed them as key terms. Instead, I alternated between using the terms *language user* or *linguistic*, which does not privilege the spoken modality as the general modality, and listing out both *speaker and signer* or both *speech and sign*, depending on the context. This included renaming 6 Speaker Demographic to 6 Language User Demographic and 8 Speech Situation and Text Characteristics to 8 Linguistic Situation and Text Characteristics. I left references to text in place because text is a necessary format for digital data; even speech datasets have textual annotations and metadata.

As part of the *What* section of 5 Language Varieties, examples of language descriptions were provided for documentation authors who might not be familiar with how to communicate information about language varieties. The first example, "English as spoken in Palo Alto, California," was presented to show how geographic region information could be incorporated into the description, especially for languages with varieties spoken around the world. The second example, "Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin," was included to show how scripts and multilinguality might be described. To show that signed languages can also have these considerations, I added, "French Sign Language as used in Marseille, France" to the list of examples.

To motivate the request for demographic information in connection with language use, we included a significant explanation of the relationship between language use and identity in 6 Language User Demo-

graphic, complete with references to the linguistics literature. For each instance of a reference for spoken languages, I added a reference to the same linguistic finding for signed languages. For the finding that linguistic variation correlates with demographic characteristics, I cited Kusters and Lucas (2022), who present the sociolinguistics of signed languages. I then cited Quinto-Pozos (2008), who described linguistic interference in the productions of signers of ASL and Mexican Sign Language (LSM), for the finding that properties of a first language affect the production of a second language. Finally, for dysarthria in signed languages, I referred to Tyrone (2014) who outlines the neurological causes and symptoms of dysarthria in signers. While these references may not be applicable to some teams who chose to use C3DAR, they still contribute to making C3DAR generally more inclusive and may help to raise awareness of and inspire interest in signed language research.

6.3.2 Changes to the Schema Elements

Changes to the schema in response to the retrospective investigation included providing new motivations for schema elements designed to address community perspectives and best practices around design and community collaboration. In several schema elements, similar motivations are listed as to why documentation readers may benefit from those elements: 1) to support assessments of the dataset match to an intended use case and 2) to support such assessments by third party users of a technology developed using the dataset. These motivations in the community context still apply, but I added an additional benefit, "to support community assessments, comparisons, and cataloging of currently available data" to 5 Language Varieties, 6 Language User Demographic, 7 Annotator Demographic, 8 Linguistic Situation and Text Characteristics and 10 Capture Quality. This new community-addressed benefit speaks to the goals of indigenous data sovereignty and signed language corpus development in valuing data for its own sake and needing infrastructure to manage increasing volumes of data. 11 Limitations also included the original two motivations, but I added a third benefit distinct from the cataloging benefit added to the other elements, "to evaluate the degree to which this dataset has contributed towards community goals." Limitations of datasets could be an important consideration for future collaborations to investigate or for community governance structures to consider when using the data. The intent for 17 Glossary was originally to provide definitions for technical terms. The ethical guidelines also emphasized the importance of using the community's vocabulary and

definitions within research and research dissemination. To support this additional function of 17 Glossary, I included the following explanation after the original motivation for documentation readers, "Definitions of local vocabulary can be important for understanding and interpreting the data in community-appropriate ways and acknowledging the validity of community knowledge and ways of knowing." These changes help to acknowledge how community documentation readers might interact with C3DAR and community datasets.

The changes across the motivations for dataset and documentation creators were more idiosyncratic. For example, in motivating why dataset creators should include 2 Executive Summary in their documentation, the original explanation listed documents that may be supported by such a summary like grant proposals and project reports. To include an example for community members I specified "project reports *to the community*" in this motivation. In 4 Documentation for Source Datasets, the motivations were simply broadened to include dataset design inspiration, in addition to inspiration for how to characterize a dataset and document important details, as a motivation for referencing source datasets. References to community ethical review processes were added to 13 Disclosure and Ethical Review with encouragement to see these processes as support structures toward ensuring the community sees tangible benefits from research: "If a community has an ethical review process, engagement with this process can help surface community-specific concerns with the dataset creation and help guide the dataset creation to support community goals." I also added to my drafted motivation for 15 Maintenance, directly addressing communities: "For communities, developing a maintenance plan may help in considering archiving options along with their benefits, risks, and costs prior to data collection." Maintenance considerations are important from a technical sense of being able to access data but also from the community perspective of being able to benefit from the data in the future and access not only the data but the community knowledge it may contain. The changes to the motivation for documentation authors in 17 Glossary again emphasize the importance of community definitions throughout the documentation, "Using local terminology throughout the documentation centers the community's understanding of the data and its cultural significance." With these additions, community motivations for documentation are made explicit throughout the schema elements.

The changes to the content of the *What* sections were modest in comparison to the *Why* sections. First I added a fourth topic to include in 2 Executive Summary, "a short description of how the community has been involved in the project." While community involvement might be apparent from the affiliations of the dataset

and documentation authors, there may be other ways in which the community want to be acknowledge for their contributions. Including a description of those contributions in 2 Executive Summary ensures that they will be seen by documentation readers, in line with the values of recognition and respect described in §5.4. To 3 Curation Rationale, I added a prompt for addressing community goals in the rationale: "How will the dataset support community goals?" Finally, in 6 Language User Demographic and 7 Annotator Demographic, the *What* section specified that demographic categories should be locally appropriate; I added to this that the definitions should be determined by the community themselves.

The new best practices align with several of the points made across the motivations and content changes. In 1 Header, the new best practice added to the discussion of community involvement in 2 Executive Summary: "Discuss with community partners how they would prefer to be acknowledged for their contributions. For some communities, coauthorship is appropriate, while others may have another preferred method. Consider also how to acknowledge contributions such as consultations on local knowledge, reviewing materials, and other efforts supporting the development of the project." Several best practices added also address community terminology and definitions. I edited the second best practice in 5 Language Varieties to suggest avoiding harmful language ideologies by using the community name and descriptions that the community recommend using. A new best practice for 8 Linguistic Situation and Text Characteristics touches on both community defined-terms and deference to the community in sharing culturally-sensitive topics: "When describing the cultural context, use community vocabulary, concepts, and interpretations to convey the cultural significance, when deemed appropriate for public dissemination by the community."

Guidance on culturally-sensitive data and protecting the community informed a number of best practices. Possible interactions with consent processes and the responsibilities of those using resulting datasets were added to a best practice for 4 Documentation for Source Datasets: "If the source dataset was collected under specific consent conditions, ensure that those conditions allow for further reuse and distribution as needed by the current dataset. When in doubt, contact the source dataset manager and ask about developing a new opt-in consent procedure for the language users who created the source data to agree to the new use and dissemination of their data." The same best practice was edited in 6 Language User Demographic and 7 Annotator Demographic. The best practice already advocated for reporting demographic information as a range to protect the people in the data from being identified; an additional suggestion advocates for

discussing appropriate demographic information with community representatives prior to collection. The only change to 9 Preprocessing and Data Formatting was to add a best practices for anonymization considerations for signed languages: "When anonymizing video or image data, modifications to the data such as blurring faces may remove necessary linguistics context and information, especially for signed languages. If language users in the dataset have not agreed to public dissemination of their video or image data without anonymization, consider all available methods for protecting the language users' privacy, such as access restrictions, and ensuring the usefulness of the dataset for the community." This best practice is intended to inform video processing as a linguistic practice so that signed language data is not rendered unintelligible by well-meaning data anonymization.

Several best practices for protecting the community were added to 14 Distribution. The first concerns determining access restrictions based on community perceptions of what data is appropriate to share and with whom: "Review the data with community representatives to determine which portions of the dataset will be culturally appropriate to share broadly, which portions should be restricted to relevant groups, and which portions should be accessible to community members only." The second best practice reminds the team to make a copy of the dataset locally available to the community in addition to the general distribution method. The third addresses licensing mechanisms and permitted uses: "When choosing terms for a license, copyright, or ToU, consider uses that will be allowed as well as uses that will be disallowed. The community should decide on whether they want to allow third-party uses such as research, use in court, technical development, and commercialization." The examples of contexts of use were drawn from examples in the licenses and guidelines analyzed in the retrospective investigation. Again alluding to community ethical review processes, the new best practice for 13 Disclosure and Ethical Review recommends reporting on any interactions with review boards and their feedback. Finally, one best practice was developed for 15 Maintenance and another was duplicated from 1 Header. The duplicated best practice is reproduced here: "Consider web accessibility and the longevity of dataset location (e.g., university archives or a community-owned repository), especially with respect to how language community members will access the data." This best practice was originally used in reference to the citation for the documentation, but also applies to the creation of the dataset. The new best practice states, "We recommend having a process for removing data, in the event that someone would like to have their data or community-sensitive data removed from the dataset."

This kind of process may be difficult to implement. A framework needs to be in place to ensure that the person requesting the removal actually has the authority to make the request (i.e., they are requesting the removal of their own data or a relative's data). Further considerations are needed for how to comply with the request if the dataset is able to be copied and downloaded locally. Despite the challenges, establishing such a process with the community is critical for respecting and supporting community accountability structures. To sum up, the changes to the schema elements resulting from the retrospective investigation largely focus on making explicit community-led definitions and decisions related to protecting the community members, protecting the community knowledge, documenting the community contributions to the dataset design and creation, and ensuring the dataset is available and useful to the community.

6.3.3 General Best Practices for Collaboration

As discussed in §5.4, many of the values expressed by communities with regards to ethical research focused on care and relationship building. Project support of these values may be evidenced in the problem framing if designed to address a community-defined goal, in the methods chosen to collect the data if using community-informed methods, and in the use of community terms throughout the documentation. But upholding these values also entails many actions that are not visible from the perspective of documentation, mainly in the day-to-day interactions between development team members and between the developers and the community. Wanting to convey these best practices but not having an appropriate location in the DSV2 schema, I created a new set of GBPs, separate from the original GBP. I called these new GBPs *general best practices for collaboration* and renamed the original GBPs *general best practices for documentation*. The general best practices for collaboration (GBPC) are as follows.

The first GBPC advocates for empathy for all stakeholders and awareness for how hierarchies might develop within the team.

> 1. Collaboration requires honesty, respect and care for team members, the community, and the community's history and values (South African San Institute, 2017; Hudson et al., 2010). Be mindful of asymmetrical power relations throughout the project within this historical context and how these interact with community cultural norms (Harris et al., 2009; Ontario Federation of Indigenous Friendship Centres, 2016).

Many of the indigenous guidelines explicitly highlighted honesty, respect, and care as important values in research (e.g. Hudson et al., 2010; South African San Institute, 2017; Ontario Federation of Indigenous Friendship Centres, 2016). Both Harris et al. (2009) and Ontario Federation of Indigenous Friendship Centres (2016) advocate for mutual respect in the context of historical oppression and its implications in present-day hierarchies of power.

The second GBPC concerns language use among the project collaborators.

> 2. Whenever possible, communicate in the language the community prefers. Relying on interpretation services may affect the results of the project and the research team's ability to understand the community's perspective, so it is best if at least one team member is able to communicate in the language of the community.

Signed language guidelines especially emphasize the necessity of working in the signed language of the community as much as possible. Even writing, as discussed in §2.2.2, disenfranchises deaf collaborators and shifts understandings of the research into spoken-language perspectives.

Centering the community is discussed in several of the new schema element best practices, but it also helps to frame the broader picture of the collaboration. GBPC 3 advocates for centering the community by prioritizing community decisions and understandings in the research design and analysis.

> 3. The project should center the needs and understandings of the community. The community should be involved in determining the project goals, methods, and evaluation criteria. The community's knowledge and ways of knowing are valid without reaffirmation via mainstream understandings and analysis (Ontario Federation of Indigenous Friendship Centres, 2016).

Western science has long held other ways of knowing in contempt and required "proof" that community knowledge was legitimate through Western methodologies. Drawn from Ontario Federation of Indigenous Friendship Centres (2016), this GBPC asserts that community knowledge and community researchers do not need to be legitimized according to the terms of Western science.

Community research may learn from the experiences of community members or establish community boards to facilitate communication between the research team and the community. In either case, the community members selected will have an impact on the research. GBPC 4 encourages the research team to

consider community diversity when learning from community members and transparently communicating these considerations to the community during the research.

> 4. Be aware of the relevant axes of diversity within the community. Community representatives should reflect the community diversity, which may be uniquely defined depending on what demographic information and personal characteristics are most salient to the community. Be transparent in any recruiting processes. Recruitment of particular community members should be transparent as to why those community members were selected for the role so as not to create mistrust from the community with respect to the project or negatively impact the recruited community member (World Federation of the Deaf Expert Group on Developing Countries, 2016).

World Federation of the Deaf Expert Group on Developing Countries (2016) in particular stressed transparency in recruitment processes and the potential harms of recruiting community members without publicly disclosing the reasons for the recruitment.

While academic processes may be more or less efficient, they are usually sensitive to the pace of academia. Community processes have their own expected timelines that research collaborators should be aware of. This is captured in GBPC 5:

> 5. Allow time for negotiation processes according to community customs as well as for feedback and reviewing processes. While community collaborators may be aware of this difference, communicating about expected time frames of both academic and community processes can help the project members to prepare ahead of time for setbacks or find other ways to make use of time spent waiting (Coeur d'Alene Tribe of Idaho and University of Idaho, 2015).

Coeur d'Alene Tribe of Idaho and University of Idaho (2015) especially emphasized that researchers and academic institutions should show care and patience for differing timelines as a way to respect indigenous communities' tribal sovereignty and established governance processes.

One topic that may require significant negotiation time is how benefits of the research will be returned to the community and in what form. GBPC 6 encourages conversations around benefits and highlights commercialization as a topic that the research team and the community should agree upon.

6. Discuss the benefits that all parties will derive from the dataset, related projects, and the collaboration itself. In particular, the outcomes should include tangible and meaningful benefits to the community that address their self-identified needs. Consider whether commercialization will be allowed on the dataset or products derived from the dataset, and if so, how the benefits and responsibilities of commercialization will be managed (Argumedo et al., 2011).

Commercializing research can cause conflicts because of inequalities as a result of capitalism and complicated legal structures around ownership and rights when money is involved. Discussing whether commercialization will even be a consideration can inform later decisions in the dataset design.

The next GBPC asserts the collaboration team's responsibility in protecting both community members and community knowledge.

7. The community and its knowledge should be protected against risks related to the project or resulting from later use of the dataset. The community members may need to be protected from physical and psychological harm, disparagement or disrespect, and confidentiality breaches. Community knowledge may be deemed sensitive and therefore inappropriate to include in any publications or publicly distributed data. Discuss the possible risks and develop mitigation strategies with the community. The implication is not that the collaboration team should be able to foresee all harms, but rather that active measures should be put in place to prevent harms and assess risks.

Despite academic ethical practices mandating that researchers "do no harm," definitions of harm are commonly defined by the researchers rather than the community at risk of being harmed. Engaging with communities in open discussions of risks helps outside researchers understand the harms that the community members may be most concerned about. Furthermore, collaboratively developing mitigation strategies empowers the community to propose solutions that address their concerns and may be integrated with established community processes.

Consent processes are established practice in academia, but culturally-appropriate consent may require additional preparations. GBPC 8 urges the research team to design their consent processes in collaboration with the community.

8. The community should be involved in developing culturally appropriate procedures for on-going free, prior, informed and educated consent. This may include the language(s) that the procedure will be available in, whether a written version will be available, and how to make the project methods, potential risks and benefits, and confidentiality procedures clear to the community. Avoid assuming that potential benefits are obvious, making exaggerated claims, and understating the potential risks. The community should decide whether this consent is individual or collective.

The language that the consent is communicated in is critical to informed and educated consent, in line with GBPC 2. The GBPC for consent also points to the possibility that collective consent may be required, depending on the situation.

In order for free, prior, informed, educated consent to be given, the research team must ensure that community members understand how the collected data may be used, who will own that data, if it will be shared in some form, and, if so, how it will be shared. GBPC 9 reminds the research team that the community retains ownership of the community knowledge shared during the research.

9. Ownership of the community's knowledge, cultural heritage, and data belongs with the community. Copies of the dataset and any other products should be returned to the community physically and/or in an accessible format. Discuss the ownership and management of the project deliverables and document the terms in schema element 14 Distribution.

This GBPC advocates for early planning of data distribution and the return of data and other artifacts to the community. It points to schema element 14 Distribution to support the discussion with further considerations around the ownership of other project outputs and specific best practices.

GBPC 10 encourages the research team to meet with some frequency in order to establish trust and build relationships. However conflicts can happen in any collaboration; planning ahead for such an event can help make the situation easier to resolve.

10. Plan to meet periodically with collaborators to discuss updates and relevant questions. Establish a mediation process for handling disagreements as they arise.

Although this consideration is not related to data, it helps the research by encouraging sustainable and accountable collaboration practices.

Even without outside accountability structures, researchers can still be accountable to the community by sharing the progress of the research as it moves forward. GBPC 11 recommends community accountability in forms that are most accessible to the general community.

> 11. Share updates with the community in a way that is transparent and comprehensible to those outside the project.

The final GBPC echoes the best practice in 1 Header with regards to acknowledgement:

> 12. Each member of the project team should receive acknowledgement and due credit for their contributions to the project in a way that is meaningful to the team member.

While the best practice in 1 Header advocates for acknowledgement across many different kinds of community contributions, this best practice concerns the collaboration team members specifically. Academic publications may be appropriate acknowledge for some collaborators, but others may not receive benefits from that method. The wording leaves the options open to the individuals to consider.

These general best practices for collaboration are intended to help frame expectations at the start of collaborative projects. These expectations involve the collaboration's responsibilities to the community at large and the responsibilities that the project team members have to each other. They aim to help start conversations around difficult topics like oppression and discrimination and make assertions for community-led research.

6.4 Summary

In this technical investigation, I drew from three sources to develop the first version of C3DAR. The first is data statements Version 2 (DSV2) as presented in Chapter 3. This provides the foundation from which I build C3DAR. Repurposing the DSV2 schema from a retrospective documentation toolkit intended for documenting datasets that have already had their data collected to a dataset design and planning toolkit required shifting the perspective of the text in the toolkit with respect to the dataset being documented (§6.1). Following these changes to the exising text, I again used datasheets to add new schema elements for dataset

maintenance and distribution (§6.2). Finally in §6.3 I applied the lessons learned from the retrospective investigation in Chapter 5 to the working version of C3DAR. This included using more inclusive linguistic terminology for signed languages, updating best practices and schema elements, and adding a set of general best practices for collaboration. These efforts all contribute to the making of C3DAR Version 1.

Chapter 7

Discussion

In this chapter, I present a discussion of the potential uses of C3DAR. I use value scenarios to imagine possible use cases (Nathan et al., 2007; Czeskis et al., 2010) and analyze their implications (§7.1). These value scenarios also help to surface values, including those supported by C3DAR and those that are in tension with C3DAR or with each other. The tensions, further discussed in (§7.2), point to the limitations of C3DAR in its development and in its ability to support community resistance to structural oppression. More broadly, the limitations of C3DAR reflect the general limitations of documentation and best practices as tools for accountability and transparency without infrastructures and enforcement mechanisms to motivate and maintain structural change in how data is treated. I discuss some limitations in §7.3 and pose open questions for future work.

7.1 Value Scenarios

As discussed in §2.4, value scenarios are a method in the value sensitive design methodological toolbox for imagining possible outcomes of using a technology and the long-term or widespread impact on stakeholders (Nathan et al., 2007; Czeskis et al., 2010). Bender and Friedman (2018) use value scenarios to surface values and value tensions in imagining the uptake of their initial proposal for the data statements Version 1 schema. In the first scenario, a medical NLP system for tagging social media data includes a subcomponent that, unknown to the developers of the system, is only trained on US and UK English data. The tagging system is adopted in a region that uses a different local variety of English and performs poorly; due to the lack of

documentation accompanying the system, it takes a long investigation to uncover the source of the problem and the system ends up impeding the local healthcare system. The second scenario considers the widespread uptake of data statements in NLP publications. This uptake leads to the identification of gaps in the collective language dataset catalog, which researchers are now able to address. Finally, the third scenario imagines the impacts to dataset publication and authorship diversity if data statements were to be standardized and made a requirement to publish as academic conferences. Using these scenarios, they highlight the possible benefits of data statements in identifying mismatches between datasets and potential use cases and more information on which languages see more dataset creation. They also highlight a possible value tension between inclusion and standardization. To preemptively mitigate these risks, Bender and Friedman (2018) suggest further investigations into the use of data statements and encourage establishing mentoring programs for writing data statements.

Following Bender and Friedman (2018), I use three value scenarios to imagine possible outcomes and impacts on stakeholders should C3DAR see uptake. The first imagines the use of C3DAR in the context of an indigenous language reclamation project and a university archive collaborating on the digitization and repatriation of written documents in the indigenous language. The second imagines C3DAR as used by a deaf developer and a linguist collaborating on open-source software and data for an app that works with ASL. The third scenario imagines what it might look like if C3DAR becomes a required document for collaborations seeking IRB approval and how automation to support the process might alter C3DAR and impact collaborations. I provide a brief analysis to sum up each scenario.

7.1.1 Language Reclamation Scenario

In this scenario, an indigenous community working on a language reclamation project contacts a university archive holding documents written in their language for help. They use C3DAR to plan out the digitization of the documents, the management of the data, and the rep/matriation of the original documents.

Scenario An indigenous community located in the United States decides to contact their nearest university to collaborate on teaching materials. The university has an archive that contains documents in the community's language, but the documents have yet to be digitized or translated. Having heard of C3DAR, the university archivist suggests using C3DAR to plan the digitization process. Initially the archivist assumes

that the resulting data and teaching materials will be made publicly available using open licenses to protect the community's rights to the data, as many previous digitization projects have done. However, while discussing the Distribution schema element during the design phase of the project, the community representatives reject this idea, stating that portions of the community data must be restricted and can only be shared with other community members. With this understanding, the archivist and community representatives are able to develop a restricted-access system for the data and a process for providing access to community members.

Discussions for the digitization of the documents and possible ways to incorporate the data into the community's language learning curriculum go smoothly. A community elder is able to help editing with a final translation of the documents after the community members create initial drafts. With the data and the translation digitized, the community and the archivist are able to negotiate with the university for the rep/matriation of the original documents in exchange for research access to the unrestricted portions of the digitized translation dataset for community-approved researchers. The citation and documentation to the dataset recognize the equal contributions of the community members and the archivist.

Analysis With the help of a clear outline for what goes into the dataset development, archiving, and distribution process, the community members and the archivist are able to work through assumptions that may have otherwise harmed the community and build a trusting relationship during the project. The community members take the lead in decisions regarding the appropriate translations and what the most useful teaching applications would be, and the archivist's technical skills and efforts to help the community understand and build their own skills with the data infrastructure are appreciated. The mutual respect developed during the project provides a foundation for future collaborations between the community and other researchers in which both the community and the researchers benefit.

7.1.2 ASL Personal Assistant Scenario

As a second scenario, a deaf software developer comes up with an idea for an app that uses ASL to input commands into a smart device. The developer is aware of varieties of ASL but unsure of how to handle it, so they contact a linguist at a local university for help. The linguist suggests the two of them use C3DAR to devise a data collection strategy. The best practices provided in C3DAR point to considerations for privacy

that the developer was previously unfamiliar with. These considerations help the collaborative team to create an innovative idea.

Disclaimer: I, a hearing person, came up with this scenario in order to set up an imaginary use case for C3DAR. I have not consulted with deaf individuals about this idea and I am not suggesting that this kind of technology is necessarily what deaf communities would want to see built.

Scenario In the not too distant future, smart technology in the home is affordable and allows modifications in the form of open-source software. A philanthropic deaf innovator comes up with an idea for a free app: a kitchen assistant that takes as input ASL instead of speech or text. The ASL-based app is installed on smart refrigerators that have screens and built-in external cameras on the doors. The basic functions of the app are simple tools to assist in the kitchen, for example starting a timer that flashes when the time is up or adding ingredients to a digital grocery list.

The innovator, however, is aware of the regional variation in ASL, so they contact a deaf university to discuss collaboration and responsible collection of representative data. A linguist responds with interest and suggests they use C3DAR to plan the data collection. As they discuss things like user privacy, anonymization processes, and open-data management, the two decide not to go forward with the data collection plan as they agree that the risk to individuals in the dataset is too great. Instead, the developer and the linguist come up with an algorithm that allows a person to locally train the app to replace the preinstalled ASL commands (the developer's data) with their own varieties of ASL.

Analysis In this scenario, a dataset is ultimately not produced from the collaboration, but C3DAR has still helped the development team make informed decisions about their product and supported the generation of an innovative idea that respects user privacy. Had the team committed to collecting the data and publicly releasing it, significant time an effort would have gone into anonymizing the data, reducing the amount of key linguistic information from facial expressions that the algorithm could use. With such as small team, the data collected would have been somewhat representative, but not exhaustive, and the tool would still have not worked for many people using different varieties of ASL. The localized algorithm instead keeps the data contained to one personal device, obviating the need for anonymization, and making the algorithm work better for any individual willing to create their own data.

7.1.3 Automation, Time-Sensitivity, and Review Processes

A third scenario imagines the effectiveness of C3DAR following its integration into an IRB process. Efforts to streamline the process for applicants leads to the programs for automatically detecting information in the executive summary and curation rationale that could be used in the other elements, helping the average collaboration project produce their documentation faster but also sometimes adding friction between collaborators.

Scenario A university IRB has decided to require C3DAR for research involving local communities. To support this requirement, the university has implemented automated processes to support faster completion of C3DAR for a given project. A community and researcher team decide to work together, so they complete the C3DAR documentation using the automation program. When the IRB approves the research, the researcher wants to start collecting data right away. However, the community representatives want more time to discuss how the research will be integrated into the community in tangible ways. This process ends up taking several weeks, leaving the researcher frustrated and the community feeling like the researcher does not respect their processes. Because of this, the collaboration is ultimately abandoned by both parties.

Analysis Because of the impersonal automation and the focus on C3DAR as a completed documentation project rather than a design process, the researcher and the community miss out on valuable opportunities to discuss each party's priorities in the project. The researcher and the community both had expectations that were not communicated prior to significant time and effort being devoted to the project. With the lack of communication and faltering trust, research that might have benefited both is given up because of competing pressures, that of the academic notion of 'publish or perish' and that of the responsibilities of the community members on the team to do their due diligence and ensure appropriate benefits from the research for their community besides the immediate research products.

7.1.4 Value Scenario Summary

I used three value scenarios to explore hypothetical use cases and the values of stakeholder groups in those contexts. The first presented a possible scenario in which C3DAR supported respect and empathy between collaborators and efforts towards negotiating mutual benefits. The second scenario imagined how C3DAR

might help collaborators consider ethical considerations of data such as privacy. Although the dataset creation was ultimately abandoned, the C3DAR toolkit still supported that collaborative decision-making process. Finally, in third scenario, the distance between the collaborators created in part due to the automated version of C3DAR produced a failure mode in which communication between the collaborators breaks down. These scenarios help to surface value tensions in collaborations and dataset creation.

7.2 Surfaced Value Tensions

The questions of benefits hinted at in the first scenario highlight the difficulty with the concept of ownership. Bragg et al. (2021) outline the various ways that ownership can be understood such as legal ownership, physical ownership, monetary ownership, and cultural ownership. While many Western understandings of ownership focus on who has the legal rights to knowledge, who physically holds data, or who may sell the research results, communities tend to understand their claims to language knowledge and data as a collective, cultural ownership. Other understandings of ownership may yet still exist for many of the world's communities. However, leveraging cultural or other forms of ownership in support of community goals for their knowledge, language, and data is a challenge as they have no strict definitions of means of enforcement. On the one hand, Western legal mechanisms such as licenses exist to ensure community rights to and benefits from data and their traditional knowledge are defined within the Western understandings of ownership and therefore protected after longstanding histories of exploitation. On the other hand, those legal mechanisms are entrenched in those same histories of capitalism and oppression. Much in the same way research is in the process of being reimagined to address these histories, what other systems and technologies could be reimagined to center other understandings and cultural values around ownership?

While negotiating ownership and benefits is always tricky, the scenario of indigenous language reclamation surfaces tensions between reciprocity and acknowledgement in the context of an indigenous and Western academic collaboration. Academics are incentivized to publish papers in academic journals (often in English); co-authorship provides an incentive for academics to collaborate on projects, and often academics co-author multiple papers together. The lead author might change depending on who devotes more time to the writing or if the paper is targeting one author's research interests. However academic publications often are not considered directly beneficial to communities. Communities may agree to authorship as

a form of acknowledgement for their contributions to the work, but tangible benefits to a community's social or economic situation are often expected in addition to appropriate acknowledgement. Acknowledgement is not equivalent to reciprocity in the context of community research as it often is in purely academic contexts.

Open data can also be implicated in misunderstandings around reciprocity and mutual benefit. The FAIR principles (Wilkinson et al., 2016) advocate for open data, leaning on the assumptions that open data is more accessible, that scientific knowledge is universally valued, and that data should be reused. These assumptions are not always true though. Open data still requires computing infrastructure to access and technical knowledge to understand; scientific knowledge may be less preferred than traditional knowledge; and some data that has been acquired without consent and through unethical and actively harmful methods should not be reused. The CARE principles (Carroll et al., 2020) help to bridge some of these assumptions with respect to ethics and community control of data, but principles can only do so much to support data being universally beneficial, if such a thing is possible and desired. The economic and environmental costs of the physical aspects of data and digital infrastructure, its *materiality* as Borning et al. (2020) call them, are often left unaddressed especially in conversations about the accessibility of digital information.

Efficiency and the time to required to build relationships may also create value tensions. The desire for efficiency creates a pull towards automation to reduce effort spent on the time-consuming aspects of research, but automation needs to be built with care to not remove the human aspects of knowledge creation and deeper understanding. Automation has the tendency to flatten and reduce the complexity of human experiences through pre-defined categories. Even something as simple as a list of languages can hide the wide range of varieties within a language and make sharing that information impossible to users of the technology. Time and empathy are needed to build understandings of differences and foster connections between collaborators and between the community; relationships cannot be automated or made more efficient. As discussed in §3.7, much time and discussion was spent working through disciplinary misunderstandings even between collaborators from relatively similar contexts and backgrounds (as academics all in technical fields in the United States). Diving into other disciplines' literature is likewise a time-consuming effort but it is a useful skill to build and collaborations can facilitate interdisciplinary learning.

While the time requirements of collaborations should not be taken lightly, collaborators' time for other occupations and responsibilities should also be respected. Efficiency with respect to the toolkit should

therefore be focused on making sure the collaborations are supported in spending time on the most important topics and are not distracted with additional or unnecessary details. The C3DAR toolkit itself is long and starting on it may feel daunting, in which case the C3DAR toolkit itself could be a force against starting a collaboration by pushing researchers to create the dataset on their own. Balancing the considerations for time, documentation, and collaboration will be a key consideration for ensuring C3DAR addresses the problems it was intended to.

7.3 Limitations

Disseminating research and data work in collaboration with communities can build trust and develop relationships (Dirks and Wanda, 2021), but documentation in and of itself does not solve problems of bias in technology nor empower communities against systems of oppression. This section considers the limitations of my methods in developing C3DAR and the limitations of documentation for the purposes of accountability and transparency. The limitations of this work are rooted in its monolinguality and my focus on English for the literature I relied on and for the language of publication for C3DAR. Somewhat ironically, some of these limitations may have been addressed by developing C3DAR, a collaborative toolkit, through the process of a collaboration but I was unable to establish more collaborative methods in the time I had available for my dissertation.

7.3.1 Methodological Limitations

In Chapter 4 and in Chapter 5, I briefly discussed the affects of focusing on English-language documents on the sample of guidelines and licenses that I analyzed in the retrospective investigation. Communities may publish their ethical guidelines in languages other than English, especially in regions where English is not the language of broader communication. My retrospective investigation is therefore systematically missing perspectives from those regions. Communities may also choose not to publish guidelines at all. As Hayward et al. (2021) point out, "Further exploration into informal, oral, and/or non-publicly facing Indigenous research ethical protocols and frameworks would be required to create a more comprehensive picture of how Indigenous peoples are defining ethical research." Such an investigation would require more careful methodology to present results while protecting sensitive community knowledge than was required

for my analysis of public-facing documents.

In addition to only analyzing English-language documents, this publication is in English and the toolkit is presented in English. This is only a reflection of my own language familiarity rather than a methodological choice. Collaboration with other dataset documentation experts will be needed to support adapting C3DAR to any other publication language. The written format of C3DAR is also firmly in the tradition of standardized writing (see Chapter 2 for discussions of other perspectives on standardized writing systems). The C3DAR toolkit using writing also invites questions about what documentation not based in writing would look like. How could the toolkit be changed to support documentation in oral languages or languages without a standard writing system? How could the toolkit be adapted to video to support documentation in signed languages? I leave these for future work for address.

7.3.2 Limitations of Best Practices

Relying on best practices has its own limitations. Researchers have to be aware of best practices in addition to having the time, resources, and inclination to follow them. Best practices may also be selectively upheld in that a researcher may choose to follow some best practices but not others. Community members and researchers may also have opposing interpretations of best practices, leading to conflict rather than collaboration. In the worst case, researchers may be aware of and discuss best practices when forming a project with academic colleagues or community members in order to convince the community to consent to a project, but forgo implementing the best practices during the project implementation.

Morton Ninomiya and Pollock (2017) provide examples of ethical dilemmas in reconciling community-based indigenous research and academic practices that are generally not covered in existing ethical best practices. These examples were drawn from a case study involving an indigenous community and a university-affiliated health study. Morton Ninomiya and Pollock describe three dilemmas that they argue could not be addressed using best practices. In the first, a lack of communication between community leaders and university organizers resulted in two studies with overlapping goals occurring simultaneously, causing confusion for the community members who participated in both studies and for the researchers who were unsure as to how to share data across the studies to reduce the burden for community participants. In the second, dissemination of the project results caused stress for the researchers who were uncertain as to what findings should

be shared with the community and how best to do so; furthermore, the need for academic publication of the results placed additional responsibilities on the community members who could review and potentially co-author academic papers. Finally, the third dilemma involved the mismatches between the university budget policies and community practices with regards to tracking funds. The possible solutions to these practical issues are not directly apparent from ethical best practices, however Morton Ninomiya and Pollock present reasonable solutions for the problems and each solution is accompanied by active communication between the relevant parties to resolve the matter. While values contained in ethical best practices do not provide the answers, having empathy and treating people with respect can help the process of finding possible paths forward in many cases. Where empathy and respect are lacking, however, the weaknesses of principles are again exposed; without accountability systems supporting those principles, there is "no mechanism to consistently enforce or hold researchers and their institutions accountable to Indigenous communities" (Morton Ninomiya and Pollock, 2017, pg. 32).

Researcher accountability can also be necessary in cases of good intentions on the part of the researcher, as time constraints or simple forgetfulness prevent research results and data from being returned to the community once a project is complete. As much as data has the potential to be beneficial to communities if applied towards community goals, data stuck on a hard drive or lost within gigabytes of university data is of no use to the community. In addition to encouraging best practices and open communication with communities, universities and institutions could enforce requirements for setting up back-up systems for accessing university data in cases where a previous point of contact for the community leaves the institution or becomes unavailable for some amount of time. More practical accountability measures and fail-safes for data could help prevent "files, libraries, and archives from becoming cemeteries for Indigenous languages" (Gaby and Woods, 2020, pg. e277).

7.3.3 Limitations of Documentation as a Bias Mitigation Tool

There are a number of reasons why documentation may be considered an attractive tool for addressing bias and transparency concerns in ML. First, the tool fits in line with academic practices where written documents are first-class artifacts. In this way, it is an immediately actionable toolkit for academics that results in a concrete product. Those products can be tied to existing reward systems in academic contexts;

they can be included in papers submitted to conferences and they can contribute to more citations as a dataset is easier to use when documented. Documentation also supports reproducibility and helps to avoid time and effort being used to make up for poor system documentation (Sambasivan et al., 2021). Sambasivan et al. (2021) surveyed ML projects across India, East and West African countries, and the USA for negative downstream impacts as a result of data issues, such as time and effort lost in using incorrect data, barriers to completing models, and data that had to be abandoned due to it no longer being usable. They found that insufficient documentation contributed to negative downstream impacts in 20% of the 53 cases and call for further incentives to reward proper data documentation.

As Gansky and McDonald (2022) rightly point out, however, documentation suffers from the same weakness as best practices, in that there are no accountability structures to enforce documentation or any further actions to remedy harms once documentation has been published. Economic and political structures outside of academia are not incentivized to be more transparent and less biased, even with newer laws like the European Union's General Data Protection Regulation (GDPR). Gansky and McDonald argue instead for research efforts toward the following two goals:

> "(1) how to adapt the deployment of digital systems to proactively address common fairness, accountability, and transparency issues — and design mechanisms for contextually appropriate participation by rightsholders in each stage of ongoing system evolution; and (2) methods for establishing and facilitating interoperability with external, independent governance institutions and infrastructure as a functional requirement for deployment of data-centric systems" (Gansky and McDonald, 2022, pg. 1990).

While C3DAR generally aligns with the first goal argued for by Gansky and McDonald, C3DAR frames the involvement of community rightsholders as collaboration rather than participation. This framing emphasizes the community's role as equal partners in the design of their datasets and technology built on their data. However, Gansky and McDonald's (2022) second goal appeals to government institutions as the arbiters of fairness and is less applicable in the community context; government institutions have been just as complicit in rights violations against minoritized communities as companies and universities. A reinterpretation of their argument to empower tribal governance structures would be more appropriate, but the judicial route has been more complex for deaf communities whose rights to signed language are intertwined with disability

rights. Miceli et al. (2022) similarly ask for more from documentation research: "More than diagnosing 'the source of bias,' documentation should aim at interrogating work practices and decision-making hierarchies within and among organizations" by acknowledging the collaborative contributions to data production by data workers in crowd-sourcing and annotation contexts (pg. 9). While this perspective on data creation considers more hierarchical structures for data teams than the (ideally) horizontal structure of equal partners in community collaborations, the suggestions to interrogate data creation practices and the histories behind data categories certainly apply to the C3DAR context.

Critical work is often dismissed for not presenting an alternate solution, but the point of these critiques is not for one person or small group to have the "right answer." Instead, critical work acknowledges how current systems are failing to support people and provides opportunities for more discussion and new paths forward. It also makes space for learning about how people are directly and indirectly impacted by technology and by systems, imagining wholly different ways of engaging with research, with technology and with one another that are not based in the current and historic systems of oppression.

7.4 Summary

In this chapter, I presented value scenarios for imagining future uses of C3DAR. These value scenarios in §7.1 provide insight into what stakeholder groups and what contexts to continue investigating in. They also surfaced value tensions in collaborative work and in C3DAR itself (§7.2). In addition to tensions, I also discuss the limitations of this dissertation and the limitations in the framing of best practices and documentation as ethical within systems that are often not ethical (§7.3). Critiques of documentation and best practices point to the lack of enforcement structures within society at large, but enforcement structures within society at large have often failed to work for minoritized language communities or actively caused more harm. Accountability structures for community-researcher collaborations developed in line with the principles of restorative justice may begin to repair these harms, provided that the accountability structures support community self-determination, guarantee that community knowledge, artifacts, and research results are returned to the community, and ensure that researchers understand the harm that has been done as a result of research that dispossesses communities of their knowledge and agency. Further critical work is needed to find localized paths forward for these communities. In the next chapter, I present a future empirical

investigation for engaging with community researchers.

Chapter 8

Toward an Empirical Investigation

With the C3DAR Version 1 toolkit in hand, next steps in the iterative design process include empirical investigations into the use of C3DAR with a variety of stakeholder groups. In this chapter, I provide an overview of a potential empirical investigation of C3DAR. I present a stakeholder analysis of community dataset documentation and potential questions to learn more about the values that these stakeholder groups find important in documentation (§8.1). In particular, I identify community researchers, that is, academics who themselves identify as members of minoritized communities, as an important initial stakeholder group to engage for future development of the project. I also propose a methodology and a set of questions for understanding to what degree C3DAR supports collaborative dataset creation and where it may be improved upon (§8.2). In §8.3, I report on some preliminary discussions with researchers currently working on language datasets for their communities. Their responses to the pilot questions and initial feedback on C3DAR point to open questions for future investigation.

8.1 Stakeholder Analysis

As discussed in §2.4, stakeholder analyses are one method for identifying whose values to consider in technological design and how their experiences with the technology might be impacted by design decisions (Friedman et al., 2006a, 2017). Stakeholders may be categorized as either *direct* stakeholders who interact directly with the C3DAR toolkit either by using it in their own collaboration or by reading the completed documentation or as *indirect* stakeholders who may never see the resulting documentation but nonetheless

are affected by others' use of it. For the purposes of measuring the impacts of C3DAR, it is also important to distinguish between stakeholders who provide perspectives from their communities and stakeholders who provide perspectives from the outside societal or academic context. Still further there are stakeholders who navigate between the community contexts and outside contexts and can provide perspectives from both contexts.

To identify stakeholder groups, I refer again to the guidelines and licenses collected in Chapter 5. These documents address many localized groups of stakeholders for indigenous communities. For example, the *Protocol and Best Practice for the Research on and Public Distribution of Information from Projects involving Indigenous Peoples* provides direct advice to investigators, administrators, and data stewards working with the Schitsu'umsh (Coeur d'Alene) tribe (Coeur d'Alene Tribe of Idaho and University of Idaho, 2015). Depending on their positionality, these groups may be outside the community or consider themselves as intermediaries between their communities and the broader societal institutions they are affiliated with like universities and archives (discussed further below). *The Standard of Conduct for Research in Northern Barkley and Clayoquot Sound Communities* also references how research projects may need to interact with other research and government organizations like national park managers (Clayoquot Alliance for Research, Education and Training, 2005). These kinds of organizations may be involved in collaborations or read instances of C3DAR, or they may be impacted by the datasets created with C3DAR. *The Community Biocultural Protocols of the Potato Park Communities in Peru* and *The Khoikhoi Peoples' Rooibos Biocultural Community Protocol* address inter-tribal structures as well as intra-tribal structures in laying out the benefit sharing mechanisms used by their communities (Argumedo et al., 2011; National Khoisan Council & Cederberg Belt Indigenous Farmers Representatives, 2019). These community structures might directly use C3DAR to manage new data project, to understand a technology developed with a dataset documented using C3DAR, or to inform new policies related to data and research benefits. In their paper *Working towards ethical guidelines for research involving the Sámi*, Holmberg (2021) point to the many Sámi community representatives in national and international government organizations across the Nordic countries. Community representatives may read C3DAR in order to understand interactions between policy and technology and as part of efforts to build community data infrastructure and capabilities. The Te Ara Tika Guidelines for Māori Research Ethics give guidance to community representatives on ethics research boards and stress

the importance of respecting the roles of families responsible for protecting cultural knowledge and relatives of individuals in the dataset (Hudson et al., 2010). Families may not be involved in directly using C3DAR, but they are impacted by the datasets created with C3DAR and technologies developed with those datasets. In considering indigenous concepts of stakeholders, many indigenous communities include their natural environments and specific landscapes and resources as relations who are impacted by the transmission of traditional knowledge and stories. For example, *The Community Biocultural Protocols of the Potato Park Communities in Peru* governance structure provides space for considering the nearby mountains within the scope of the protocol as owners of the land, animals, and people following local traditions. Leonard (2021) also includes ancestors and artifacts as stakeholders to consider in indigenous language reclamation; these stakeholders would be indirectly impacted by the use of C3DAR in that C3DAR may help to continue the traditions and language use associated with those artifacts and ancestors.

Both indigenous and deaf communities greatly value their communities' being able to teach their children their language and customs. Teachers and children, both those who are learning the language now and those who will learn the language in the future are therefore stakeholders of C3DAR. Teachers may use C3DAR to understand how data and language technology might be used within their curriculum, while students would potentially benefit from innovative and culturally appropriate lessons using language technology. De Meulder (2021) also points to interpreters as potential indirect stakeholders impacted by language technology; while indigenous interpreters may be unquestioned in their roles within the community, signed language interpreters are necessarily hearing and their inclusion within the deaf community may be controversial. Table 8.1 presents these direct and indirect stakeholders of C3DAR in terms of their context.

The dual roles of community researchers are not without their complications. Conflicts between the values and needs of the community and the values and expectations of the academy can lead community researchers to feel like they have to choose one or the other due to limited time, resources, and energy (O'Brien, 2017). Community researchers have also been instrumental in pushing for incremental changes in academic perceptions of what "counts" as research and who "counts" as a researcher (Haualand, 2017; Hermes, 2012, see e.g.). Community researchers therefore have much experience related to negotiating value tensions and weighing potential outcomes in the context of community research. Investigations into how to better support community researchers might draw on Raji et al. (2021)'s notion of transversality in which

Context	Direct	Indirect
In Community	Community representatives (in gov. & review boards)	Families responsible for knowledge Individuals in the dataset
	Inter- and intra-community structures	Relatives of those in the dataset
	Language teachers	Ancestors and artifacts
	Community researchers	Children (present and future)
Both	Community data stewards	Second-language learners
		Community administrators
		Natural resources (land, water, etc.)
		Interpreters (from the community)
Outside Community	Collaborator researchers	Administrators (funding, university)
	Data stewards (librarians, archivists)	Data stewards (librarians, archivists)
	Other researchers and orgs.	Other researchers and orgs.
		Interpreters (from outside the community)

Table 8.1: Direct and indirect stakeholders of C3DAR

problems across multiple disciplines or knowledges are best addressed using methods that build across those knowledges or on the indigenous framework of Two-Eyed Seeing in which indigenous and Western methods are integrated in research (Iwama et al., 2009). Such investigations could also provide valuable lessons into the uses of C3DAR and how it might be improved.

8.2 Empirical Investigation Proposal

The retrospective investigation provided a very broad and general overview of the values that communities find the most important for research and data. To learn more about specific stakeholders and the application of their values with respect to collaborative dataset creation, I propose the following empirical investigation. In this future empirical investigation, I would seek out collaborators working with spoken indigenous languages, collaborators working with signed languages, and particularly collaborators who are themselves community members. The goal would be to learn from community researchers who hold both technical knowledge and community knowledge with respect to dataset creation to not only see both communities' and academics' perspectives, but also to understand the possible interactions between the perspectives. This dual understanding of both communities' values is critical to supporting future collaborations using C3DAR, particularly in cases where each member of the collaboration only has perspectives on one kind of knowl-

edge.

The following questions may help in learning about collaborators experiences with dataset creation and the stakeholders who they keep in mind during the process:

- What does their process for creating a dataset look like?
- Who do they work with?
- Who do they imagine will directly use their datasets?
- Who do they imagine will be indirectly affected by their datasets?
- Do they consider societies, future generations, natural resources, and/or other entities who may be impacted?
- Who do they imagine will interact with the documentation of their dataset?

As discussed in the previous section, stakeholders may be broadly interpreted and include more abstracted notions of affected parties such as future generations and natural entities and artifacts. These questions would hopefully surface some of those considerations. To explore collaborators' values in dataset creation, I propose the following questions:

- What do they think is important in dataset creation?
- What values do they try to support and how?
- Have there been times when two of these values have seemed to be in tension with one another?
- What interactions with their dataset do they actively try to prevent?
- What actions do they want their dataset documentation to support?

Additionally, specific values may be selected in order to ask the collaborators about how those values in particular manifest in their work, if those values do not come up in the conversation naturally. This discussion to understand the values that are prioritized in the collaboration helps determine the dimensions along which C3DAR must be evaluated in order to understand in what ways C3DAR is successful in supporting collaborative dataset design and in what what ways C3DAR may continue to be improved.

Given sufficient time, the empirical investigation may also invite the collaborators to use C3DAR to develop a dataset. This dataset could be real, if the collaboration was already planning to create one, or hypothetical. Following an introduction to C3DAR, collaborators could be asked for their feedback on the toolkit. Questions could include:

- Given their experiences, do they imagine they would be able to complete the C3DAR toolkit when designing a new dataset?
- How do they think the toolkit would support the values they mentioned previously?
- Which values are unsupported or less supported?

As the collaborators used the C3DAR toolkit to develop their datasets, interspersed meetings could provide information on collaborators' experiences with C3DAR over time. Meeting two more times would provide the opportunity to investigate how their initial impressions of C3DAR may differ from their impressions after using the toolkit. The meetings could happen once when they are partially through the toolkit and a second time when they are finished with it. During these meetings, the discussion would be focused on the ways that the toolkit is or is not supporting their work and ways in which it might be improved.

8.3 Preliminary Discussions

To provide an example of the kinds of findings that the proposed empirical investigation could surface, I posed the initial questions to a small number of researchers involved in community dataset creation. I contacted two researchers directly through mutual contacts as well as recruited three more by posting a request on social media (Twitter). In particular, I sought feedback from indigenous and signed language community developers. I met with a total of five experts.

I had discussions with two signed language experts, Dr. Julie A. Hochgesang of Gallaudet University in the United States and Dr. Rose Stamp of Bar-Ilan University in Israel. Dr. Hochgesang has collaborated on multiple datasets for American Sign Language (ASL, ISO 639-3 code [ase]) (Hochgesang et al., 2017-2022; Dudis et al., 2020) and studies the particular ethical concerns related to signed language datasets (Hochgesang et al., 2010; Bragg et al., 2021).[1] Dr. Stamp has worked on collaborative projects for both

[1] A colleague of Dr. Hochgesang's kindly interpreted for us during our meeting.

British Sign Language (BSL, ISO 639-3 code [bfi]) (Stamp et al., 2015) and Israeli Sign Language (ISL, ISO 639-3 code [isr]) (Stamp et al., 2022; Morgan et al., 2022).

I also met with three experts working with spoken languages: Dr. Benjamin E. Frey, Dr. Santiago Esteban, and Merve Ünlü Menevşe. Dr. Frey of the University of North Carolina Asheville has collaborated with NLP experts on a machine translation system for English (ISO 639-3 code [eng]) and Eastern Cherokee (ISO 639-3 code for Cherokee [chr]) (Zhang et al., 2020, 2021) and has been working on developing more Eastern Cherokee language learning applications using NLP techniques (Frey, 2020; Zhang et al., 2022). Dr. Santiago Esteban works with Argentinian Spanish (ISO 639-3 code [spa]) datasets from electronic health records in the context of the government healthcare system of Buenos Aires, Argentina (Esteban et al., 2016; Volij and Esteban, 2020, e.g.). Merve Ünlü Menevşe is a Ph.D. student at Boğaziçi University in Turkey working on NLP appliations in the Turkish language (ISO 639-3 code [tur]) (Ünlü Menevşe et al., 2022).

I met with each researcher in a virtual meeting at least once. One researcher had time for a second meeting, and two more provided feedback on the C3DAR toolkit asynchronously. To start the discussions, I introduced the motivations for C3DAR and my research in dataset documentation. I then asked each researcher about their experiences with dataset creation and collaboration.

Though they all worked in very different contexts, each researcher brought up a sense of responsibility they felt in dataset creation, either to the community they worked with or to the quality of the work. They discussed the pressures that worked against collaboration, such as time constraints, a lack of support in their institution for long-term data work, or interdisciplinary misunderstandings. Funding was frequently a concern or even a barrier to starting a project.

The feedback for C3DAR generally considered it to be a useful tool (though, without having yet used it). One researcher expressed interest in using the toolkit for their own work. More support was requested for specific suggestions, like using accessibility tools for blind and low-vision readers and understanding how to select appropriate licensing terms for data. One researcher put forward that the documentation could potentially add to bureaucratic workloads and be seen as a checklist to complete prior to research even being able to start. Both of these ideas present avenues for future investigation. One investigation could explore the ways that licenses are selected and present case studies of datasets with different terms of use.

Another investigation could look into where the balancing point for documentation lands in terms of

making reasonable expectations of researchers and imposing too much on their time. An initial survey could ask researchers how much time they generally spend on project design and documentation. One factor to consider is that the time spent on project design and documentation may change when the project is carried out by a team with a single affiliation as opposed to when the project is carried out in collaborations between research institutions and communities. An empirical study in which researchers designed a real or hypothetical dataset using C3DAR could then compare the time spent on the project design and documentation in the study with the time that researchers reported in the initial survey. Whether the teams are able to complete the documentation or not would be an indication that the documentation may be too time-consuming.

8.4 Summary

Investigations with stakeholder groups such as the one suggested in this chapter will help to ensure that C3DAR continues to be adapted to stakeholder needs. These investigations, however, will be scoped by investigators to address particular questions or explore stakeholder needs in identified conditions, naturally limiting the range of possible findings. These will provide insight into predefined use cases such as with communities just starting to create datasets for their languages. However, it may be that C3DAR could also be used by communities with existing datasets that need documentation, similar to the retrospective use case of data statements, or used by archives to normalize documentation across many repositories. Further unanticipated use cases will likely appear as C3DAR is applied to varied contexts.

Chapter 9

Conclusion

In this work, I present developments for one documentation toolkit (data statements) and propose a second documentation toolkit – Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources (C3DAR) – consisting of a specification and a set of best practices for planning the creation of language datasets with communities grounded in collaborative and community-led research paradigms. This chapter presents my concluding thoughts, including a summary of my investigations (§9.1), potential applications of C3DAR, and suggestions for future investigations (§9.3). Throughout this dissertation, I explored two questions: How can we develop tools and practices for dataset documentation that are responsive to the needs of more varied stakeholders groups? How can documentation support collaborative language technology work between researchers and language communities?

To prepare to address these questions, I first conducted a literature review surveying relevant research from technical communities and two minoritized language communities. While not comprehensive, Chapter 2 outlined the ethical concerns of signed language communities and indigenous language communities in the context of historical oppression and naive applications of technological solutions. It also presented efforts towards ethical technological development within the fields of natural language processing (NLP) and machine learning (ML). Finally, I introduced value sensitive design and its role in the development of data statements for NLP, a documentation schema for presenting information about language datasets (Bender and Friedman, 2018). The introduction to value sensitive design methodologies also serves to contextualize my use of them throughout my investigations.

9.1 Summary of Investigations

Chapter 3 presented my co-authored work on the development of data statements Version 2. The investigations we conducted towards Version 2 were centered on understanding documentation practices for technical communities and how to support those communities in writing data statements. To learn about how NLP and ML practitioners might use data statements and to collect feedback on best practices for writing data statements, we organized a three-day workshop as an empirical investigation. During this workshop, participants drafted their own data statements and shared their experience of the process with us. After analyzing the participant feedback, we drafted an interim version of data statements including participant feedback and recommendations as new schema elements and best practices. We then conducted a technical investigation in which we compared the interim version of data statements with datasheets for datasets (Gebru et al., 2021), another documentation format developed with similar intended purposes and format as data statements. This comparison yield final changes to the data statement schema elements, resulting in a completed data statement Version 2 schema (Appendix A) and an accompanying guide for writing data statements.

To continue my investigations, I centered the perspectives of a different set of stakeholders, minoritized language communities. I set out to develop a new toolkit for facilitating community-research collaborations and supporting minoritized language communities in maintaining control over their data and knowledge systems in the context of language dataset creation. Chapter 4 presented my methodology towards this new toolkit, again using value sensitive design methods to lay out investigations.

I conducted two investigations to develop C3DAR: a retrospective technical investigation and a technical investigation. In Chapter 5, I reported on the results of my retrospective technical investigation in which I analyzed ethical guidelines and licenses written primarily by both indigenous and signed language communities from around the world. Coding these guidelines and licenses for the values they expressed revealed patterns in prioritized values across geographies, modalities, and authoring institutions. Despite comparisons with a second annotator showing a high degree in variation across annotators, the coding exercise in this investigation surfaced not only the frequencies of different values, but also the different actions that communities point to as evidence of support for those values in research.

Chapter 6 detailed my efforts to integrate into one new collaboration and dataset design toolkit the content and style of data statements Version 2, additional content from datasheets, and the lessons from the

retrospective investigation. Using data statements Version 2 as the foundation of the new toolkit, I added two new schema elements from datasheets, Maintenance and Distribution. The questions in these datasheets sections were edited for style and to address a broader audience than the original intended audience of datasheets. I then leveraged the lessons from the guidelines and licenses to develop a new set of general best practices for collaboration and further best practices across the schema elements, old and new. Final edits to address the intended language communities audiences concluded the development of C3DAR.

9.2 Potential Uses and Limitations

I discussed the limitations of this work in Chapter 7. This discussion ranged from the specific limitations of the investigations conducted and the general limitations of documentation and ethical guidelines in guiding ethical research. In the retrospective investigation, the method of artificially grouping interdependent constellations of values into rigid, distinct categories proved to be highly subjective. Still further, collecting only English-language guidelines and licenses excluded all considerations from communities where English is not considered the language of broader communication. While the toolkit encourages publication and communication in the language of the community, it still imposes on communities the need for *written* documentation. Further investigation is needed to understand how to present dataset documentation in audio (for oral languages) and video (for signed languages) formats. Documentation remains a limited tool for shifting power; its widespread use without accompanying actions to support community goals risks increasing the frequency of performative ethics and adding bureaucratic barriers to research with communities.

To explore potential futures in which C3DAR is adopted by various communities, I presented three value scenarios (Nathan et al., 2007; Czeskis et al., 2010). Analyzing these hypothetical scenarios helped to surface tensions for C3DAR in the context of an indigenous reclamation project, an imagined ASL-based software product, and as required documentation for a review process. These scenarios imagine possibilities, but they are not predictive nor do they replace learning from uses across localized contexts. To begin to understand use in context, empirical investigations are needed.

9.3 Towards Future Investigations

Chapter 8 presents a pilot empirical investigation and some initial discussions with community researchers. Following a stakeholder analysis of the direct and indirect users of C3DAR, I suggest community researchers hold important insight for further development of C3DAR as individuals who are familiar with both the pressures of being a scholar in academia and the pressures of being a community member with responsibilities and personal ties to their community. I present a set of pilot questions to understand community researchers values and considered stakeholders in their work. I then share initial responses from community researchers who were willing to contribute their experiences in developing language datasets for their communities.

In order to instantiate the findings from the empirical investigation in the C3DAR toolkit, further technical investigations would need to be conducted. One possibility would be to conduct case studies with community researchers. These case studies could be used to investigate how C3DAR supports the development of a new community dataset and where it could be improved to better support the community's prioritized values. Similarities and differences across these case studies could then be used to make general improvements to C3DAR or provide more insight in how value tensions arise and are addressed in the case studies.

Another avenue for investigation could be to adapt the Diverse Voices methodology for engaging with experiential experts (Young et al., 2019). Young et al. (2019) developed the method in order to collect feedback on tech policy drafts from experiential experts who may be impacted by the policy if implemented. It may be that the method could also be used to collect feedback from experiential experts such as community representatives and administrators on drafted instances of C3DAR for specific datasets. The questions posed could investigate the participants' experiences in reading the drafted C3DAR documentation and how they may be impacted on technology built using the documented dataset or it could investigate how C3DAR might be more broadly used to support their needs.

Further empirical studies using varied methods could explore questions such as: How does C3DAR facilitate communities and researchers in building new collaborations without prior relationships? How does C3DAR support collaborations in negotiating tensions and disagreements in the dataset design? How could C3DAR be used in community research review processes? How might C3DAR be adapted to and published in other languages and modalities? These questions and more will help to improve existing mechanisms and

frameworks for supporting communities in maintaining agency over their data and incrementally develop new toolkits as appropriate for the community context.

9.4 Conclusion

The main contribution of this dissertation is a toolkit called Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources (C3DAR), intended to support the collaborative design of language community datasets. The development of this toolkit was grounded in investigations of the values that are important to minoritized language communities. While each community faces its own unique challenges, many communities resonate with calls for community self-determination and community-led research directions. C3DAR asserts communities' rights to design their research goals and control their data. Along the way, it provides best practices for navigating collaborations and fostering relationships between people who come from unique disciplinary and community backgrounds.

Bibliography

Gilles Adda, Khalid Choukri, Irmgarda Kasinskaite, Joseph Mariani, Hélène Mazo, and Sakti Sakriani, editors. 2019. *Proceedings of the 1st International Conference on Language Technologies for All*. European Language Resources Association (ELRA), Paris, France.

Alfonso Aguavil Calazacón, Catalina Calazacón Calazacón, Ramón Aguavil Calazacón, Juan Aguavil Calazacón, Milton Calazacón Calazacón, José Jacinto Aguavil Loche, Francisco Aguavil Alopi, Alejandrino Aguavil Aguavil, Rosa Aguavil, Manuel Aguavil Calazacón, Domingo Zaracay, Primitivo Aguavil, and Eloy Alopi Aguavil. 1983-2012. Tsafiki Documentation Project. DoBeS Archive MPI Nijmegen. Accessed on 9 Mar 2023 at https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0001_46BA_3.

Pius Akumbu. 2014. Multimedia Documentation of Babanki Ritual Speech. Endangered Languages Archive. Accessed on 9 Mar 2023 at http://hdl.handle.net/2196/00-0000-0000-0002-EB3C-5.

Alaska Native Knowledge Network. 2006. Alaska Federation of Natives Guidelines for Research. Accessed on 9 Mar 2023 at http://www.ankn.uaf.edu/IKS/afnguide.html.

Americans with Disabilities Act. 1990. S.933 - 101st Congress (1989-1990): Americans with Disabilities Act of 1990. https://www.congress.gov/bill/101st-congress/senate-bill/933.

Jonathan D. Amith. 2020. Catalogue of Nahuatl (Glottolog = high1278; ISO 639-3 = azz) recordings and transcriptions from the municipality of Cuetzalan del Progreso, Puebla, Mexico. Endangered Languages

Archive. Accessed on 8 Mar 2023 at https://www.elararchive.org/uncategorized/IO_e53e81ef-5dc5-4cf0-bbf5-03babaef80bb/.

Hannah Anglin-Jaffe. 2015. De-Colonizing Deaf Education: An Analysis of the Claims and Implications of the Application of Post-Colonial Theory to Deaf Education. In K. Lesnik-Oberstein, editor, *Rethinking Disability Theory and Practice*. Palgrave Macmillan UK, United Kingdom.

Archive of the Indigenous Languages of Latin America. 2017a. Access. Accessed on 9 Mar 2023 at https://ailla.utexas.org/site/depositors/access.

Archive of the Indigenous Languages of Latin America. 2017b. AILLA Conditions of Use. Accessed on 9 Mar 2023 at https://ailla.utexas.org/site/rights/use_conditions.

Archive of the Indigenous Languages of Latin America. 2017c. AILLA License. Accessed on 9 Mar 2023 at https://ailla.utexas.org/site/rights/license.

Alejandro Argumedo, Asociación ANDES (Peru), the Potato Park Communities, and International Institute for Environment and Development. 2011. Community Biocultural Protocols: Building Mechanisms for Access and Benefit Sharing among the Communities of the Potato Park based on Customary Quechua Norms. Technical report, International Institute for Environment and Development. Accessed on 9 Mar 2023 at https://www.iied.org/g03168.

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John Richards, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13.

Asia Pacific Forum of National Human Rights Institutions and Office of the High Commissioner for Human Rights. 2013. United Nations Declaration on the Rights of Indigenous Peoples: A Manual for National Human Rights Institutions. Accessed on 22 Jun 2023 at https://www.ohchr.org/sites/default/files/Documents/Issues/IPeoples/UNDRIPManualForNHRIs.pdf.

John Bandy and Nicholas Vincent. 2021. Addressing "Documentation Debt"' in Machine Learning: A Retrospective Datasheet for BookCorpus. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. SIGCIS Conference http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf.

Sarah C. E. Batterbury. 2012. Language justice for Sign Language Peoples: the UN Convention on the Rights of Persons with Disabilities. *Language policy*, 11(3):253–272.

Sarah C. E. Batterbury, Paddy Ladd, and Mike Gulliver. 2007. Sign language peoples as indigenous minorities: Implications for research and policy. *Environment and planning A*, 39(12):2899–2915.

Seyma Bayram and Rebecca Hersher. 2023. A new satellite could help clean up the air in America's most polluted neighborhoods. *National Public Radio*. Accessed on 21 Jun 2023 at https://www.npr.org/2023/06/19/1179670466/air-pollution-satellite-baltimore-climate-change.

Maurício C.A Belo, John Bowden, John Hajek, Nikolaus P. Himmelman, and Alexandre V. Tilman. 2002-2006. DoBeS Waima'a Documentation. DoBeS Archive MPI Nijmegen. Accessed on 9 Mar 2023 at https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0008_3938_D.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Batya Friedman, and Angelina McMillan-Major. 2021a. A guide for writing data statements for natural language processing. Available at http://techpolicylab.uw.edu/data-statements/.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Erik Blasch, James Sung, Tao Nguyen, Chandra P. Daniel, and Alisa P. Mason. 2019. Artificial Intelligence Strategies for National Security and Safety Standards. *arXiv preprint arXiv:1911.05727*. `https://doi.org/10.48550/arXiv.1911.05727`.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tracey A. Bone, Erin Wilkinson, Danielle Ferndale, and Rodney Adams. 2022. Indigenous and Deaf People and the Implications of Ongoing Practices of Colonization: A Comparison of Australia and Canada. *Humanity & society*, 46(3):495–521.

Alan Borning, Batya Friedman, and Nick Logler. 2020. The 'Invisible' Materiality of Information Technology. *Communications of the ACM*, 63(6):57–64.

danah boyd and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Karen L. Boyd. 2021. Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2).

Annelies Braffort. 2002. Research on computer science and sign language: Ethical aspects. In *Gesture and Sign Language in Human-Computer Interaction*, pages 1–8, Berlin, Heidelberg. Springer Berlin Heidelberg.

Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective. *ACM Transactions on Accessible Computing*, 14(2).

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. Exploring Collection of Sign Language Datasets: Privacy, Participation, and Model Performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA. Association for Computing Machinery.

Ben Braithwaite. 2020. Ideologies of linguistic research on small sign languages in the global South: A Caribbean perspective. *Language & Communication*, 74:182–194.

E Bruijn and Gail Whiteman. 2010. That which doesn't break us: Identity work by local 'stakeholders'. *Journal of business ethics*, 96(3):479–496.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334):183–186.

Deborah Cameron, Elizabeth Frazer, Penelope Harvey, MBH Rampton, and Kay Richardson. 1992. *Researching Language: Issues of Power and Method*. Politics of language. Routledge.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*.

Stephanie Russo Carroll, Desi Rodriguez-Lonebear, and Andrew Martinez. 2019. Indigenous Data Governance: Strategies from United States Native Nations. *Data Science Journal*, 18:31.

Anne H. Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, 96(4):e200–e235.

Bagele Chilisa. 2020. *Indigenous research methodologies*, 2nd edition. SAGE, Thousand Oaks, California.

Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*.

Kimberly Christen. 2015. Tribal archives, traditional knowledge, and local contexts: Why the "s" matters. *Journal of Western Archives*, 6(1). 3.

Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. 2012. A comparative study of adaptive, automatic recognition of disordered speech. In *Proc. Interspeech 2012*, pages 1776–1779.

Clayoquot Alliance for Research, Education and Training. 2005. Standard of conduct for research in Northern Barkley and Clayoquot Sound communities. Accessed on 8 Mar 2023 at `https://achh.ca/wp-content/uploads/2018/07/Protocol_Northern-Barkley-and-Clayoquot-Sound.pdf`.

Goedele A. M. De Clerck and Sam Lutalo-Kiingi. 2018. Ethical and methodological responses to risks in fieldwork with deaf Ugandans. *Contemporary Social Science*, 13(3-4):372–385.

Coeur d'Alene Tribe of Idaho and University of Idaho. 2015. Protocol and Best Practice for the Research on and Public Distribution of Information from Projects involving Indigenous Peoples. Accessed on 8 Mar 2023 at `https://www.sqigwts.org/sites/default/files/Protocol%20Final%20Sqigwts%2010-30-15_0.pdf`.

Donavyn Coffey. 2021. Maori are trying to save their language from big tech. *WIRED*. Accessed 20 Jun 2023 at `https://www.wired.co.uk/article/maori-language-tech`.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Richard Cokart, Trude Schermer, Corrie Tijsseling, and Eva Westerhoff. 2019. In Pursuit of Legal Recognition of the Sign Language of the Netherlands. In Maartje De Meulder, Joseph J Murray, and Rachel L McKee, editors, *The legal recognition of sign languages: Advocacy and outcomes around the world*, page 161. Multilingual Matters.

Pusch Commey. 2003. The new scramble for Africa: Biopiracy. *New African (London. 1978)*, (424).

Audrey C. Cooper and Trần Thủy Tiên Nguyễn. 2015. Signed Language Community-Researcher Collaboration in Việt Nam: Challenging Language Ideologies, Creating Social Change. *Journal of Linguistic Anthropology*, 25(2):105–127.

Juliet M. Corbin and Anselm L. Strauss. 2008. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 3rd edition. SAGE, Los Angeles, California.

Alain Couillault, Karen Fort, Gilles Adda, and Hugues De Mazancourt. 2014. Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter. In *International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

Glen Sean Coulthard. 2014. *Red Skin, White Masks : Rejecting the Colonial Politics of Recognition*. Indigenous Americas. University of Minnesota Press, Minneapolis.

Fiona Cram. 2001. *Rangahau Māori: Tona Tika, Tona Pono*, pages 35–52. Longman, Auckland.

Fiona Cram. 2006. Talking Ourselves UP. *AlterNative: An International Journal of Indigenous Peoples*, 2(1):28–43.

Onno Crasborn, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anne de Meijer, and Anna Sáfár. 2006-2017. Corpus NGT. DoBeS Archive MPI Nijmegen. Accessed on 8 Mar 2023 at `https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0021_6AC3_1`.

Kate Crawford. 2017. The Trouble with Bias. Keynote at NeurIPS `https://www.youtube.com/watch?v=fMym_BKWQzk;%20neurIPS%20keynote`.

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation*, 3(1).

Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. 2010. Parenting from the Pocket: Value Tensions and Technical Directions for Secure and Private Parent-Teen Mobile Safety. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA. Association for Computing Machinery.

Alexandra D'Arcy and Emily M. Bender. 2023. Ethics in linguistics. *Annual Review of Linguistics*, 9(1):49–69.

Dominique David-Chavez, Stephanie Russo Carroll, Serena Natonabah, Brianne Lauro, and Andrew Martinez. 2020. Supporting Indigenous data stewards: Indigenous governance and ethics in scientific research. Accessed 3 Mar 2023 at `https://indigenousdatalab.org/indigenousdatastewards/`.

Jenny L. Davis. 2017. Resisting rhetorics of language endangerment: Reclamation through Indigenous language survivance. *Language Documentation and Description*, 14:37–58.

Maartje De Meulder. 2021. Is "good enough" good enough? Ethical and responsible development of sign language technologies. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 12–22, Virtual. Association for Machine Translation in the Americas.

Maartje De Meulder, Verena Krausneker, Graham Turner, and John Bosco Conama. 2019a. *Sign Language Communities*, pages 207–232. Palgrave Macmillan UK, London.

Maartje De Meulder, Joseph J Murray, and Rachel L McKee. 2019b. Introduction: The legal recognition of sign languages: Advocacy and outcomes around the world. In Maartje De Meulder, Joseph J Murray, and Rachel L McKee, editors, *The legal recognition of sign languages: Advocacy and outcomes around the world*, pages 1–16. Multilingual Matters.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Leon Derczynski and Alex Speed Kjeldsen. 2019. Bornholmsk Natural Language Processing: Resources and Tools. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 338–344, Turku, Finland. Linköping University Electronic Press.

Aashaka Desai, Lauren Berger, Fyodor O Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, and Danielle Bragg. 2023. ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition. *arXiv preprint arXiv:2304.05934*.

Desert Knowledge Cooperative Research Centre. 2007. Desert Knowledge CRC Protocol for Aboriginal Knowledge and Intellectual Property. Accessed on 8 Mar 2023 at `http://www.nintione.com.au/resource/DKCRC-Aboriginal-Intellectual-Property-Protocol.pdf`.

Lisa G Dirks and Pratt Wanda. 2021. Technology to support collaborative dissemination of research with alaska native communities. *AMIA ... Annual Symposium proceedings*, 2021:398–407.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

P. Dudis, J.A. Hochgesang, E. Shaw, and M. Villanueva. 2020. Introduction to "Motivated Look at Indicating Verbs in ASL (MoLo)" Project. HDLS14 Poster Presentation.

Durbin Feeling Native American Languages Act. 2023. S.1402 - 117th Congress (2021-2022): Durbin Feeling Native American Languages Act of 2022. `https://www.congress.gov/bill/117th-congress/senate-bill/1402`.

Arienne M Dwyer. 2006. Ethics and practicalities of cooperative fieldwork and analysis. *Essentials of language documentation*, 178.

Vijay A. D'Souza. 2015. Documentation and description of the Hrusso Aka language of Arunachal Pradesh. Endangered Languages Archive. Accessed on 9 Mar 2023 at `http://hdl.handle.net/2196/00-0000-0000-000F-BF58-9`.

Penelope Eckert and John R. Rickford. 2001. *Style and Sociolinguistic Variation*. Cambridge University Press, Cambridge, UK; New York, NY.

R. A. R. Edwards. 2012. *Words Made Flesh: Nineteenth-Century Deaf Education and the Growth of Deaf Culture*. The history of disability. New York University Press, New York.

Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford.

Endangered Languages Archive. 2023. How to use the archive. Accessed on 9 Mar 2023 at https://www.elararchive.org/how-to-use/.

Michael Erard. 2005. How linguists and missionaries share a bible of 6,912 languages. *The New York Times*. Accessed 23 Jun 2023 at https://www.nytimes.com/2005/07/19/science/how-linguists-and-missionaries-share-a-bible-of-6912-languages.html.

Michael Erard. 2017. Why Sign-Language Gloves Don't Help Deaf People. *The Atlantic*. Accessed 21 Jun 2023 at https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/.

Joseph Errington. 2008. *Linguistics in a colonial world: A story of language, meaning, and power*. Blackwell Publishing Ltd., Malden, MA.

Santiago Esteban, Fernando Vázquez Peña, and Sergio Terrasa. 2016. Translation and cross-cultural adaptation of a standardized international questionnaire on use of alternative and complementary medicine (I-CAM-Q) for Argentina. *BMC complementary and alternative medicine*, 16(1):1–7.

Esther Martinez Native American Languages Preservation Act. 2006. H.R.4766 - 109th Congress (2005-2006): Esther Martinez Native American Languages Preservation Act of 2006. https://www.congress.gov/bill/109th-congress/house-bill/4766.

Every Student Succeeds Act. 2015. Committees - S.1177 - 114th Congress (2015-2016): Every Student Succeeds Act. https://www.congress.gov/bill/114th-congress/senate-bill/1177.

Jordan Fenlon, Kearsy Cormier, and Adam Schembri. 2015. Building bsl signbank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2):169–206.

First Nation's Information Governance Centre. 2018. The First Nations Principles of OCAP®. Available at https://fnigc.ca/ocap. Accessed 13 June 2023.

Lance Forshay, Kristi Winter, and Emily M. Bender. 2016. SignAloud Open Letter. Accessed 23 Jun 2023 at `http://faculty.washington.edu/ebender/papers/SignAloudOpenLetter.pdf`.

Jens Forster, Daniel Stein, Ellen Ormel, Onno Crasborn, and Hermann Ney. 2010. Best practice for sign language data collections regarding the needs of data-driven recognition and translation. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 92–97, Valletta, Malta. European Language Resources Association (ELRA).

Karën Fort, Gilles Adda, and K Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Karën Fort and Alain Couillault. 2016. Yes, We Care! Results of the Ethics and Natural Language Processing Surveys. In *international Language Resources and Evaluation Conference (LREC) 2016*, Proceedings of the international Language Resources and Evaluation Conference (LREC) 2016, Portoroz, Slovenia.

R. Edward Freeman. 1984. *Strategic Management: A Stakeholder Approach*. Pitman series in business and public policy. Pitman, Boston, MA.

Benjamin Frey. 2020. "Data is nice:" Theoretical and pedagogical implications of an Eastern Cherokee corpus. In Wilson de Lima Silva and Katherine J. Riestenberg, editors, *Collaborative Approaches to the Challenges of Language Documentation and Conservation: Selected papers from the 2018 Symposium on American Indian Languages (SAIL)*, number 20 in Language Documentation & Conservation Special Publication, pages 38–53. University of Hawai'i Press.

Batya Friedman, Edward Felten, and Lynette I Millett. 2000. Informed consent online: A conceptual model and design principles. *University of Washington Computer Science & Engineering Technical Report 00–12–2*, 8.

Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, Cambridge.

Batya Friedman, David G. Hendry, and Alan Borning. 2017. *A Survey of Value Sensitive Design Methods*, volume 11 of *Foundations and trends in human-computer interaction*. Now Publishers, Boston - Delft.

Batya Friedman, Daniel Howe, and Edward Felten. 2002. Informed Consent in the Mozilla Browser: Implementing Value Sensitive Design. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, volume 8 of *HICSS '02*, page 247, USA. IEEE Computer Society.

Batya Friedman, Peter H. Kahn Jr., Jennifer Hagman, Rachel L. Severson, and Brian Gill. 2006a. The Watcher and the Watched: Social Judgments About Privacy in a Public Place. *Human–Computer Interaction*, 21(2):235–272.

Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3):330–347.

Batya Friedman, Jr. Peter H. Kahn, and Alan Borning. 2006b. Value sensitive design and information systems. In Ping Zhang and Dennis F. Galletta, editors, *Human-Computer Interaction in Management Information Systems: Foundations*, pages 348–372. M. E. Sharpe, Armonk NY.

Batya Friedman, Ian Smith, Peter H. Kahn, Sunny Consolvo, and Jaina Selawski. 2006c. Development of a Privacy Addendum for Open Source Licenses: Value Sensitive Design in Industry. In *Proceedings of the 8th International Conference on Ubiquitous Computing*, UbiComp'06, page 194–211, Berlin, Heidelberg. Springer-Verlag.

Alice Gaby and Lesley Woods. 2020. Toward linguistic justice for Indigenous people: A response to Charity Hudley, Mallinson, and Bucholtz. *Language*, 96(4):e268–e280.

Natacha Gagné. 2015. Brave New Words: The Complexities and Possibilities of an "Indigenous" Identity in French Polynesia and New Caledonia. *The Contemporary Pacific*, (2):371–402.

Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT Undermines Its Organizing Principles. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1982–1992, New York, NY, USA. Association for Computing Machinery.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010v1.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for datasets. *CoRR*, abs/1803.09010v7.

Ann E Geers, Christine M Mitchell, Andrea Warner-Czyz, Nae-Yuh Wang, Laurie S Eisenberg, CDaCI Investigative Team, et al. 2017. Early sign language exposure and cochlear implantation benefits. *Pediatrics*, 140(1).

Lily George, Juan Taori, and Lindsey Te Ata o Tu MacDonald, editors. 2020. *Indigenous Research Ethics: Claiming Research Sovereignty Beyond Deficit and the Colonial Legacy*, volume 6 of *Advances in Research Ethics and Integrity*. Emerald Publishing, United Kingdom.

Gustavo Godoy and André Sanches de Abreu. 2022. Yman har ma'e pandu ha: Myths and accompanying co-speech gestures in Ka'apor. Endangered Languages Archive. Accessed on 9 Mar 2023 at `https://www.elararchive.org/dk0704`.

Ben Green. 2019. "Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, pages 1–7.

Colette Grinevald. 2003. Speakers and documentation of endangered languages. *Language documentation and description*, 1:52–72.

Ken Hale, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C. England. 1992. Endangered languages. *Language*, 68(1):1–42.

Thomas Hanke and Jordan Fenlon. 2022. Creating Corpora: Data Collection. In Jordan Fenlon and Julie A. Hochgesang, editors, *Signed Language Corpora*, The Sociolinguistics in Deaf Communities series, Volume 25, pages 18–45. Gallaudet University Press, Washington, DC.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in Size and Depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological*

Challenges and Application Perspectives, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Raychelle Harris, Heidi M Holmes, and Donna M Mertens. 2009. Research Ethics in Sign Language Communities. *Sign language studies*, 9(2):104–131.

Hilde Haualand. 2017. When Inclusion Excludes. Deaf, Researcher–Either, None, or Both. In Annelies Kusters, Maartje De Meulder, and Dai O'Brien, editors, *Innovations in Deaf Studies: The Role of Deaf Scholars*, Perspectives on Deafness, pages 317–338. Oxford University Press, New York, NY.

Ashley Hayward, Erynne Sjoblom, Stephanie Sinclair, and Jaime Cidro. 2021. A New Era of Indigenous Research: Community-based Indigenous Research Ethics Protocols in Canada. *Journal of Empirical Research on Human Research Ethics*, 16(4):403–417. PMID: 34106784.

David G Hendry, Batya Friedman, and Stephanie Ballard. 2021. Value sensitive design as a formative framework. *Ethics and Information Technology*, 23:39–44.

Mary Hermes. 2012. Indigenous Language Revitalization and Documentation in the United States: Collaboration Despite Colonialism. *Language and Linguistics Compass*, 6(3):131–142.

Jane H. Hill. 2002. "Expert Rhetorics" in Advocacy for Endangered Languages: Who Is Listening, and What Do They Hear? *Journal of Linguistic Anthropology*, 12(2):119–133.

Joseph Hill. 2020. Do deaf communities actually want sign language gloves? *Nature Electronics*, 3(9):512–513.

Joseph C. Hill. 2017. The Importance of the Sociohistorical Context in Sociolinguistics: The Case of Black ASL. *Sign Language Studies*, 18(1):41–57.

Michael Hind, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, and Kush R Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv preprint arXiv:1808.07261*, 18:2813–2869.

Julie A. Hochgesang. 2015. Ethics of researching signed languages: The case of Kenyan Sign Language

(KSL). *Signed Languages in Sub-Saharan Africa: Politics, citizenship and shared experiences of difference*, pages 9–28.

Julie A. Hochgesang. 2022. Managing Sign Language Acquisition Video Data: A Personal Journey in the Organization and Representation of Signed Data. In *The Open Handbook of Linguistic Data Management*. The MIT Press.

Julie A. Hochgesang, Onno Crasborn, and Diane Lillo-Martin. 2017-2022. ASL Signbank. `https://aslsignbank.haskins.yale.edu/`.

Julie A. Hochgesang, Ryan Lepic, and Emily Shaw. 2023. W(h)ither the ASL corpus?: Considering trends in signed corpus development. In *Advances in Sign Language Corpus Linguistics*, pages 287–308. John Benjamins.

Julie A. Hochgesang and Nick Palfreyman. 2022. Signed Language Corpora and the Ethics of Working With Signed Language Communities. In Jordan Fenlon and Julie A. Hochgesang, editors, *Signed Language Corpora*, The Sociolinguistics in Deaf Communities series, Volume 25, pages 158–195. Gallaudet University Press, Washington, DC.

Julie A. Hochgesang, Pedro Pascual-Villanueva, Gaurav Mathur, and Diane Lillo-Martin. 2010. Building a Database while Considering Research Ethics in Sign Language Communities. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 112–115, Valletta, Malta. European Language Resources Association (ELRA).

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv preprint arXiv:1805.03677*.

Áslat Holmberg. 2021. Working towards ethical guidelines for research involving the sámi. Accessed on 8 Mar 2023 at `https://www.saamicouncil.net/documentarchive/working-towards-ethical-guidelines-for-research-involving-the-smi`.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, page 1–16. Association for Computing Machinery, New York, NY, USA.

Gary Holton, Wesley Y. Leonard, and Peter L. Pulsifer. 2022. Indigenous Peoples, Ethics, and Linguistic Data. In *The Open Handbook of Linguistic Data Management*. The MIT Press.

Lynn Hou, Ryan Lepic, and Erin Wilkinson. 2020. Working with asl internet data. *Sign Language Studies*, 21(1):32–67.

Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach, editors. 2017. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Maui Hudson, Moe Milne, Paul Reynolds, Khyla Russell, and Barry Smith. 2010. Te Ara Tika Guidelines for Māori research ethics: A framework for researchers and ethics committee members. Accessed on 8 Mar 2023 at `https://www.hrc.govt.nz/sites/default/files/2019-06/Resource%20Library%20PDF%20-%20Te%20Ara%20Tika%20Guidelines%20for%20Maori%20Research%20Ethics_0.pdf`.

Tom Humphries. 1975. Audism: The making of a word. Unpublished essay.

Tom Humphries, Poorna Kushalnagar, Gaurav Mathur, Donna Jo Napoli, Carol Padden, and Christian Rathmann. 2014. Ensuring language acquisition for deaf children: What linguists can do. *Language*, 90(2):e31–e52.

Tom Humphries, Raja Kushalnagar, Gaurav Mathur, Donna Jo Napoli, Carol Padden, Christian Rathmann, and Scott Smith. 2013. The Right to Language. *The Journal of Law, Medicine & Ethics*, 41(4):872–884.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Sarah Louise Hyett, Chelsea Gabel, Stacey Marjerrison, and Lisa Schwartz. 2019. Deficit-Based Indigenous Health Research and the Stereotyping of Indigenous Peoples. *Canadian Journal of Bioethics / Revue canadienne de bioéthique*, 2(2):102–109.

Intelligent Transportation Systems Committee of the IEEE Vehicular Technology Society and Standing Committee for Standards or the IEEE Robotics and Automation Society. 2021. IEEE Standard for Transparency of Autonomous Systems. Technical Report IEEE Std 7001-2021, Institute of Electrical and Electronics Engineers, New York, NY. `https://standards.ieee.org/ieee/7001/6929/`.

International Society of Ethnobiology. 2006 with 2008 additions. International Society of Ethnobiology Code of Ethics. Retrieved from http://ethnobiology.net/code-of-ethics/.

Amy Isard. 2020. Approaches to the anonymisation of sign language corpora. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 95–100, Marseille, France. European Language Resources Association (ELRA).

Marilyn Iwama, Murdena Marshall, Albert Marshall, and Cheryl Bartlett. 2009. Two-eyed seeing and the language of healing in community-based research. *Canadian journal of native education*, 32(2):3–23.

Trevor Johnston. 2006. W(h)ither the Deaf Community? Population, Genetics, and the Future of Australian Sign Language. *Sign Language Studies*, 6(2):137–173.

Trevor Johnston. 2008. Auslan Corpus. Endangered Languages Archive. Accessed on 8 Mar 2023 at https://www.elararchive.org/dk0001.

Trevor Johnston. 2009. Creating a corpus of Auslan within an Australian national corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*.

Trevor Johnston, Adam Schembri, Robert Adam, Jemina Napier, Darlene Thorton, Julia Allen, Karin Banna, Donovan Cresdee, Louise de Beuzeville, Lindsay Ferrara, Dani Fried, Della Goswell, Michael Gray, Ben Hatchard, Gabrielle Hodge, Gerry Shearim, Jane van Roekel, Lori Whynot, and Steve Cassidy. 2023. Auslan Signbank Acknowledgements. Accessed on 8 Mar 2023 at https://auslan.org.au/about/acknowledgements/.

Joint Technical Committee ISO/IEC JTC 1, Information Technology, Subcommittee SC 38, Cloud Computing and Distributed Platforms. 2020. Cloud computing and distributed platforms Data flow, data categories and data use — Part 1: Fundamentals. Technical Report ISO/IEC 19944-1:2020(en), International Organization for Standardization and the International Electrotechnical Commission, Geneva. https://www.iso.org/standard/79573.html.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ka Haka 'Ula O Ke'elikolani College of Hawaiian Language. 2019. Kani'aina Conditions of Access. Accessed on 8 Mar 2023 at https://ulukau.org/kaniaina/?a=p&p=conditionsofaccess&e=-------en-20--1--txt-tpIN%7ctpTI%7ctpTA%7ctpCO%7ctpTY%7ctpLA%7ctpPR%7ctpSG%7ctpTO%7ctpTG%7ctpSM%7ctpTR%7ctpET%7ctpHT%7ctpDT%7ctpMG%7ctpSS%7ctpCL%7ctpLO----------.

Kaipuleohone. 2008. Kaipuleohone Conditions of Access. Accessed on 9 Mar 2023 at http://ling.hawaii.edu/wp-content/uploads/Kaipuleohone-ConditionsOfAccess.pdf.

Kaipuleohone. 2015. Kaipuleohone Deposit Form. Accessed on 9 Mar 2023 at http://ling.hawaii.edu/wp-content/uploads/Kaipuleohone-DepositForm.pdf.

Larry Lindsey Kauanoe Kimura (depositor). 2018. Ka Leo Hawai'i Collection. Kaipuleohone. Accessed on 8 Mar 2023 at https://scholarspace.manoa.hawaii.edu/collections/4d30e25d-c3c8-4bcf-9386-559de553ecdd?cp.page=3.

Verna J. Kirkness and Ray Barnhardt. 1991. First nations and higher education: The four r's — respect, relevance, reciprocity, responsibility. *Journal of American Indian Education*, 30(3):1–15.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. License Conditions for Movies, Pictures, Annotations, and Metadata on this Site. Accessed on 8 Mar 2023 at https://www.sign-lang.uni-hamburg.de/meinedgs/ling/license_en.html.

Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache/MY DGS – annotated. Public Corpus of German Sign Language. Version 3.0.

Verena Krausneker. 2015. Ideologies and attitudes toward sign languages: An approximation. *Sign Language Studies*, 15(4):411–431.

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

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine

Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Paul V. Kroskrity. 2005. *Language Ideologies*, chapter 22. John Wiley & Sons, Ltd.

Nicole S. Kuhn, Myra Parker, and Clarita Lefthand-Begay. 2020. Indigenous Research Ethics Requirements: An Examination of Six Tribal Institutional Review Board Applications and Processes in the United States. *Journal of empirical research on human research ethics: JERHRE*, 15(4):279–291.

Tahu Kukutai and Maggie Walter. 2015. Recognition and indigenizing official statistics: Reflections from Aotearoa New Zealand and Australia. *Statistical Journal of the IAOS*, 31(2):317–326.

Annelies Kusters, Maartje De Meulder, and Dai O'Brien. 2017a. Innovations in Deaf Studies: Critically Mapping the Field. In Annelies Kusters, Maartje De Meulder, and Dai O'Brien, editors, *Innovations in Deaf Studies: The Role of Deaf Scholars*, Perspectives on Deafness, pages 1–56. Oxford University Press, New York, NY.

Annelies Kusters, Maartje De Meulder, and Dai O'Brien, editors. 2017b. *Innovations in Deaf Studies: The Role of Deaf Scholars*. Perspectives on Deafness. Oxford University Press, New York, NY.

Annelies Kusters and Ceil Lucas. 2022. Emergence and evolutions: Introducing sign language sociolinguistics. *Journal of Sociolinguistics*, 26(1):84–98.

William Labov. 1966. *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington, DC.

Paddy Ladd. 2003. *Understanding Deaf Culture: In Search of Deafhood*. Multilingual Matters, Clevedon, England.

Paddy Ladd. 2007. Cochlear implantation, colonialism, and Deaf rights. In Linda Komesaroff, editor, *Surgical consent: Bioethics and cochlear implantation*, pages 1–29. Gallaudet University Press.

Peter Ladefoged. 1992. Another view of endangered languages. *Language*, 68(4):809–811.

Harlan Lane. 1984. *When the Mind Hears: A History of the Deaf*. Random House, New York.

Harlan Lane. 1992. *The mask of benevolence: disabling the Deaf community*. Knopf, New York.

Language and Culture Archive of Ecuador. 2010a. How To Use. Accessed on 9 Mar 2023 at `https://flacso.edu.ec/lenguas-culturas/sobre-el-archivo/para-que-sirve/?lang=en`.

Language and Culture Archive of Ecuador. 2010b. Rights. Accessed on 9 Mar 2023 at `https://flacso.edu.ec/lenguas-culturas/sobre-el-archivo/derechos/?lang=en`.

Brian Larson. 2017. Gender as a Variable in Natural-Language Processing: Ethical Considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

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

Wesley Y. Leonard. 2020. Insights from Native American Studies for theorizing race and racism in linguistics (Response to Charity Hudley, Mallinson, and Bucholtz). *Language*, 96(4):e281–e291.

Wesley Y. Leonard. 2021. Centering Indigenous Ways of Knowing in Collaborative Language Work. In *Sustaining Indigenous Languages: Connecting Communities, Teachers, and Scholars*, pages 21–34. Northern Arizona University.

Wesley Y. Leonard and Erin Haynes. 2010. Making "collaboration" collaborative: An examination of perspectives that frame linguistic field research. *Language Documentation & Conservation*, 4:269–293.

Claudia Leto, Winarno S. Alamudi, Nikolaus P. Himmelmann, Jani Kuhnt-Saptodewo, Sonja Riesberg, and Hasan Basri. 2005-2010. DoBeS Totoli Documentation. DoBeS Archive MPI Nijmegen. Accessed on 9 Mar 2023 at `https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0014_C4BF_9`.

Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. Indigenous Protocol and Artificial Intelligence Position Paper. Technical report, Aboriginal Territories in Cyberspace, Honolulu, HI. https://spectrum.library.concordia.ca/id/eprint/986506/7/Indigenous_Protocol_and_AI_2020.pdf.

William D. Lewis and Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint Rumour Stance and Veracity Prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Yvonna Lincoln and Norman Denzin. 2005. The eighth and ninth moments: Qualitative research in/and the fractured future. In Yvonna Lincoln and Norman Denzin, editors, *The SAGE Handbook of Qualitative Research*, 3rd ed. edition, page 1115–26. SAGE, Thousand Oaks, California.

Linguistic Society of America. 2011. Resolution for U.S. Government Action to Support the Preservation and Revitalization of Native American Languages. Accessed on 21 Jun 2023 at https://www.linguisticsociety.org/resource/resolution-us-government-action-support-preservation-and-revitalization-native-american.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Lassana Magassa and Batya Friedman. Under review. Toward inclusive justice: Applying the Diverse Voices design method to improve the Washington State Access to Justice Technology Principles.

Lassana Magassa, Meg Young, and Batya Friedman. 2017. Diverse voices: A how-to guide for facilitating inclusiveness in tech policy. Available at `http://techpolicylab.uw.edu/project/diverse-voices/`.

Elizabeth S Mathews. 2018. *Language, Power, and Resistance: Mainstreaming Deaf Education*. Gallaudet University Press, Cambridge.

Teresa L. McCarty, Sheilah E. Nicholas, Kari A. B. Chew, Natalie G. Diaz, Wesley Y. Leonard, and Louellyn White. 2018. Hear Our Languages, Hear Our Voices: Storywork as Theory and Praxis in Indigenous-Language Reclamation. *Daedalus*, 147(2):160–172.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data statements: From technical concept to community practice. *ACM Journal of Responsible Computing*, 00:17. Just Accepted, Article 00 (May 2023).

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35.

Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP).

Lynette I. Millett, Batya Friedman, and Edward Felten. 2001. Cookies and Web Browser Design: Toward Realizing Informed Consent Online. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, page 46–52, New York, NY, USA. Association for Computing Machinery.

Yeshimabeit Milner and Amy Traub. 2021. Data Capitalism and Algorithmic Racism. Technical report, Data for Black Lives and Demos. https://www.demos.org/research/data-capitalism-and-algorithmic-racism.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Modern Language Association. 2022. Snapshot: Language study in fall 2020. volume 54, pages 6–7.

Luisa Mok and Sampsa Hyysalo. 2018. Designing for energy transition through Value Sensitive Design. *Design Studies*, 54:162–183.

Hope Morgan, Wendy Sandler, Rose Stamp, and Rama Novogrodsky. 2022. ISL-LEX v.1: An online lexical resource of Israeli Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 148–153, Marseille, France. European Language Resources Association.

Melody E. Morton Ninomiya and Nathaniel J. Pollock. 2017. Reconciling community-based Indigenous research and academic practices: Knowing principles is not always enough. *Social Science Medicine*, 172:28–36.

Joseph J. Murray. 2015. Linguistic Human Rights Discourse in Deaf Community Activism. *Sign Language Studies*, 15(4):379–410.

Joseph J Murray. 2019. American Sign Language Legislation in the USA. In Maartje De Meulder, Joseph J Murray, and Rachel L McKee, editors, *The legal recognition of sign languages: Advocacy and outcomes around the world*, pages 119–128. Multilingual Matters.

Joseph J Murray, Wyatte C Hall, and Kristin Snoddon. 2019. Education and health of children with hearing loss: the necessity of signed languages. *Bulletin of the World Health Organization*, 97(10):711–716.

Lisa P. Nathan, Batya Friedman, Predrag Klasnja, Shaun K. Kane, and Jessica K. Miller. 2008. Envisioning Systemic Effects on Persons and Society throughout Interactive System Design. In *Proceedings of the 7th ACM Conference on Designing Interactive Systems*, DIS '08, page 1–10, New York, NY, USA. Association for Computing Machinery.

Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value Scenarios: A Technique for Envisioning Systemic Effects of New Technologies. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, page 2585–2590, New York, NY, USA. Association for Computing Machinery.

National Khoisan Council & Cederberg Belt Indigenous Farmers Representatives. 2019. The Khoikhoi Peoples' Rooibos Biocultural Community Protocol. Accessed on 6 Mar 2023 at `https://naturaljustice.org/wp-content/uploads/2020/04/NJ-Rooibos-BCP-Web.pdf`.

National Research Council, Institute of Medicine, Division on Earth and Life Studies, Board on Health Promotion and Disease Prevention, Board on Radiation Effects Research, Polar Research Board, Commission on Life Sciences, Environment and Resources Commission on Geosciences, and Committee on Evaluation of 1950s Air Force Human Health Testing in Alaska Using Radioactive Iodine-131. 1996. *The Arctic Aeromedical Laboratory's Thyroid Function Study: A Radiological Risk and Ethical Analysis*. National Academies Press, Washington, D.C.

Native American Language Resource Center Act. 2023. S.989 - 117th Congress (2021-2022): Native American Language Resource Center Act of 2022. `https://www.congress.gov/bill/117th-congress/senate-bill/989`.

Native American Languages Act. 1990. Cosponsors - S.2167 - 101st Congress (1989-1990): Native American Languages Act. `https://www.congress.gov/bill/101st-congress/senate-bill/2167`.

Serena Natonabah, Brianne D. Lauro, Dominique M. David-Chavez, and Stephanie Carroll. 2020. How are we supporting Indigenous data stewards?: Aligning Indigenous and federal environmental science research ethics guidelines. In *AGU Fall Meeting Abstracts*, volume 2020, pages SY020–0010.

Mauro Nicolao, Heidi Christensen, Stuart Cunningham, Phil Green, and Thomas Hain. 2016. A Framework for Collecting Realistic Recordings of Dysarthric Speech - the homeService Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1993–1997, Portorož, Slovenia. European Language Resources Association (ELRA).

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York.

Wade W. Nobles. 1976. Extended Self: Rethinking the So-Called Negro Self-Concept. *Journal of Black Psychology*, 2(2):15–24.

Nuu-chah-nulth Tribal Council Research Ethics Committee. 2008. Protocols & Principles for Conducting Research in a Nuu-chah-nulth Context. Accessed on 8 Mar 2023 at `https://icwrn.uvic.ca/wp-content/uploads/2013/08/NTC-Protocols-and-Principles.pdf`.

Dai O'Brien. 2017. Deaf-led Deaf Studies: Using Kaupapa Maori Principles to Guide the Development of Deaf Research Practices. In Annelies Kusters, Maartje De Meulder, and Dai O'Brien, editors, *Innovations in Deaf Studies: The Role of Deaf Scholars*, Perspectives on Deafness, pages 57–76. Oxford University Press, New York, NY.

Office of the High Commissioner for Human Rights. 2006. Convention on the Rights of Persons with Disabilities. Accessed on Jun 20 2023 at `https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-persons-disabilities`.

Office of the High Commissioner for Human Rights. 2007. United Nations Declaration on the Rights of Indigenous Peoples. Accessed on 19 Jun 2023 at `https://www.ohchr.org/en/indigenous-peoples/un-declaration-rights-indigenous-peoples`.

Pamela Oliver. 2017. Race names. Available at `https://www.ssc.wisc.edu/soc/racepoliticsjustice/2017/09/16/race-names/`. Accessed 27 June 2023.

Ontario Federation of Indigenous Friendship Centres. 2016. USAI Research Framework. Accessed on 13 Mar 2023 at `https://ofifc.org/wp-content/uploads/2020/03/USAI-Research-Framework-Second-Edition.pdf`.

Adreanne Ormond, Fiona Cram, and Lyn Carter. 2006. Researching our Relations: Reflections on Ethics and Marginalisation. *AlterNative: An International Journal of Indigenous Peoples*, 2(1):174–193.

Pacific And Regional Archive for Digital Sources in Endangered Cultures. 2014. PARADISEC Deposit Form and Conditions of Access. Accessed on 9 Mar 2023 at `https://www.paradisec.org.au/PDSCdeposit.pdf`.

Carol A Padden and Tom Humphries. 1988. *Deaf in America: Voices from a culture*. Harvard University Press, Cambridge, Mass.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Michael A. Peters and Carl T. Mika. 2017. Aborigine, Indian, indigenous or first nations? *Educational Philosophy and Theory*, 49(13):1229–1234.

President's Council of Advisors on Science and Technology. 2010. Designing a digital future: Federally funded research and development in networking and information technology. Technical report, Executive Office of the President of the United States, Washington, DC. `https://www.nitrd.gov/pubs/PCAST-NITRD-report-2010.pdf`.

Soraia Prietch, J. Alfredo Sánchez, and Josefina Guerrero. 2022. A Systematic Review of User Studies as a Basis for the Design of Systems for Automatic Sign Language Processing. *ACM Transactions on Accessible Computing*, 15(4).

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA. Association for Computing Machinery.

David Quinto-Pozos. 2008. Sign Language Contact and Interference: ASL and LSM. *Language in Society*, 37(2):161–189.

Inioluwa Deborah Raji. 2020. Handle with Care: Lessons for Data Science from Black Female Scholars. *Patterns (New York, N.Y.)*, 1(8):1–3.

Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 515–525, New York, NY, USA. Association for Computing Machinery.

Inioluwa Deborah Raji and Jingying Yang. 2019. ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles. *arXiv preprint arXiv:1912.06166*.

BD Rampey, SC Faircloth, RP Whorton, and J Deaton. 2019. National indian education study 2015: A closer look. Technical Report NCES 2019-048, Institute of Education Sciences, National Center for Education Statistics, U.S. Department of Education. Washington, D.C.

BD Rampey, SC Faircloth, RP Whorton, and J Deaton. 2021. National indian education study 2019. Technical Report NCES 2021-018, Institute of Education Sciences, National Center for Education Statistics, U.S. Department of Education. Washington, D.C.

Keren Rice. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics*, 4(1-4):123–155.

John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović. 2020. A Methodology for Creating AI FactSheets. *arXiv preprint arXiv:2006.13796*.

Michael Rießler, Anna Afanasyeva, Anja Behnke, Svetlana Danilova, Andrej Dubovcev, Aleksandra Erštadt, Dorit Jackermeier, Elena Karvovskaya, Kristina Kotcheva, Jurij Kusmenko, Maryna Litvak, Sergej Nikolaev, Kateryna Olyzko, Niko Partanen, Elisabeth Scheller, Nina Šarshina, Ganna Vinogradova, Joshua Wilbur, Evgenia Zhivotova, and Nadežda Zolotuchina. 2005-2017. Kola Saami Documentation Project: Linguistic and ethnographic documentation of the endangered Kola Saami languages. DoBeS Archive

MPI Nijmegen. Accessed on 8 Mar 2023 at https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_000A_2DC3_F.

Eva-Maria Rößler, Jan David Hauck, and Warren Thompson (eds.). 2008-present. DOBES Archive for the Aché Language. DoBeS Archive MPI Nijmegen. Accessed on 9 Mar 2023 at https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_001A_7BE0_C.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Marc Schulder and Thomas Hanke. 2022. How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages. In *13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 164–173, Marseille, France. European Language Resources Association (ELRA).

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Technical Report NIST Special Publication (SP) 1270, Includes updates as of March 2022, National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1270.

Secretariat of the Permanent Forum on Indigenous Issues. 2018. Action plan for organizing the 2019 International Year of Indigenous Languages. https://en.iyil2019.org/wp-content/uploads/2018/09/N1804802.pdf.

Christin Seifert, Stefanie Scherzinger, and Lena Wiese. 2019. Towards Generating Consumer Labels for Machine Learning Models. In *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, pages 173–179.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Ac-*

countability, and Transparency, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Sign Language Linguistics Society. 2016. SLLS Ethics Statement for Sign Language Research. Accessed 10 Mar 2023 at `https://slls.eu/slls-ethics-statement/`.

Stephen C. Slota, Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. 2020. Good systems, bad data?: Interpretations of AI hype and failures. *Proceedings of the Association for Information Science and Technology*, 57(1):e275.

Linda Tuhiwai Smith. 2012. *Decolonizing Methodologies: Research and Indigenous Peoples*, second edition edition. Zed Books, London.

South African San Institute. 2017. San Code of Research Ethics. Accessed on 6 Mar 2023 at `https://www.globalcodeofconduct.org/wp-content/uploads/2018/04/San-Code-of-RESEARCH-Ethics-Booklet_English.pdf`.

Rose Stamp, Ora Ohanin, and Sara Lanesman. 2022. The corpus of Israeli Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 192–197, Marseille, France. European Language Resources Association.

Rose Stamp, Adam Schembri, Jordan Fenlon, and Ramas Rentelis. 2015. Sociolinguistic variation and change in british sign language number signs: Evidence of leveling? *Sign Language Studies*, 15(2):151–181.

William C. Stokoe, Jr. 1960. Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *The Journal of Deaf Studies and Deaf Education*, 10(1):3–37.

Julia Stoyanovich and Bill Howe. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3).

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Communications of the ACM*, 56(5):44–54.

Felix Y. B. Sze. 2014. Preliminary Documentation of Macau Sign Language. Endangered Languages Archive. Accessed on 9 Mar 2023 at `http://hdl.handle.net/2196/00-0000-0000-000F-B674-6`.

Zeerak Talat. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Kimberly TallBear. 2013. *Native American DNA: Tribal belonging and the false promise of genetic science*. University of Minnesota Press, Minneapolis, MN.

Angoua Tano. 2013. Documentation and description of a sign language in Côte d'Ivoire. Endangered Languages Archive. Accessed on 9 Mar 2023 at `http://hdl.handle.net/2196/00-0000-0000-0008-6ABC-4`.

Te Hiku Media. 2022. Kaitiakitanga License. Github. Accessed on 9 Mar 2023 at `https://github.com/TeHikuMedia/Kaitiakitanga-License`.

Technical Committee ISO/TC 46, Information and documentation, Subcommittee SC 4, Technical interoperability. 2017. Information and documentation—The Dublin Core metadata element set—Part 1: Core elements. Technical Report ISO 15836-1:2017, International Organization for Standardization, Geneva. `https://www.iso.org/standard/71339.html`.

Technical Committee ISO/TC 46, Information and documentation, Subcommittee SC 4, Technical interoperability. 2019. Information and documentation—The Dublin Core metadata element set—Part 2: DCMI Properties and classes. Technical Report ISO 15836-2:2019, International Organization for Standardization, Geneva. `https://www.iso.org/standard/71341.html`.

Bernard T Tervoort. 1953. *Structurele Analyze van Visueel Taalgebruik binnen een Groep Dove Kinderen: Structural analysis of visual language use within a group of deaf children. Deel 2. Materiaal, Registers, Enz.* North-Holland Publishing Company.

Michael Thomas. 2014. Sakun (Sukur) Language Documentation. Endangered Languages Archive. Accessed on 9 Mar 2023 at `http://hdl.handle.net/2196/00-0000-0000-0002-CF29-4`.

Adrienne Tsikewa. 2021. Reimagining the current praxis of field linguistics training: Decolonial considerations. *Language*, 97(4):e293–e319.

Ranalda L Tsosie, Anne D Grant, Jennifer Harrington, Ke Wu, Aaron Thomas, Stephan Chase, D'Shane Barnett, Salena Beaumont Hill, Annjeanette Belcourt, Blakely Brown, and Ruth Plenty Sweetgrass-She Kills. 2022. The Six Rs of Indigenous Research. *Tribal college journal of American Indian higher education*, 33(4):n4.

Eve Tuck. 2011. Rematriating curriculum studies. *Journal of Curriculum and Pedagogy*, 8(1):34–37.

Håkan Tunón, Marie Kvarnström, and Henrik Lerner. 2016. Ethical codes of conduct for research related to Indigenous peoples and local communities: core principles, challenges and opportunities. In *Ethics in Indigenous Research*, Samiska studier, pages 57–80.

Martha E. Tyrone. 2014. Sign Dysarthria: A Speech Disorder in Signed Language. In David Quinto-Pozos, editor, *Multilingual Aspects of Signed Language Communication and Disorder*, volume 11 of *Communication disorders across languages*. Multilingual Matters.

Merve Ünlü Menevşe, Yusufcan Manav, Ebru Arisoy, and Arzucan Özgür. 2022. A Framework for Automatic Generation of Spoken Question-Answering Data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4659–4666, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Camila Volij and Santiago Esteban. 2020. Development of a Systematic Text Annotation Standard to Extract Social Support Information form Electronic Medical Records. *Studies in Health Technology and Informatics*, 270:1261–1262.

Zvjezdana Vrzić, Robert Doričić, Valter Zulijani, Ivan Gabriš, Ivana Eterović, Beti Gal Jurić, Ana Stojanović, and Toni Roce. 2023. Preservation of the Vlaški and Žejanski Language. Accessed on 8 Mar 2023 at `https://www.vlaski-zejanski.com/en/o-projektu`.

Maggie Walter, Tahu Kukutai, Stephanie Russo Carroll, and Desi Rodriguez-Lonebear. 2021. *Indigenous data sovereignty and policy*. Taylor & Francis.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Gail Whiteman. 2009. All My Relations: Understanding Perceptions of Justice and Conflict between Companies and Indigenous Peoples. *Organization studies*, 30(1):101–120.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Peter Wittenburg. 2005a. DOBES Code of Conduct. Accessed on 9 Mar 2023 at `https://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf`.

Peter Wittenburg. 2005b. DOBES Data Access and Protection Rules. Accessed on 9 Mar 2023 at https://dobes.mpi.nl/ethical_legal_aspects/DOBES-access-v2.pdf.

Peter Wittenburg. 2005c. DOBES Depositor-Archivist Agreement. Accessed on 9 Mar 2023 at https://dobes.mpi.nl/ethical_legal_aspects/DOBES-daa-v1.pdf.

James Woodward. 1975. How you gonna get to heaven if you can't talk with Jesus: the educational establishment vs. the Deaf community. In *34th Annual Meeting of the Society for Applied Anthropology. Royal Tropical Institute, Amsterdam, The Netherlands*.

James Woodward and Thomas Horejes. 2016. deaf/Deaf: Origins and Usage. In *The SAGE Deaf Studies Encyclopedia*, pages 284–287. SAGE Publications.

World Federation of the Deaf. 2019. World Federation of the Deaf Charter on Sign Language Rights for All. Accessed on 16 June 2023.

World Federation of the Deaf and Finnish Association of the Deaf. 2015. Chapter 6: Best Practices and Challenges in Sign Language Work. In *Working Together: Manual for Sign Language Work within Development Cooperation*. Finnish Association of the Deaf.

World Federation of the Deaf Expert Group on Developing Countries. 2016. Best Practices and Ethics for Development Co-operation Projects. Accessed on 6 Mar 2023 at https://wfdeaf.org/news/resources/may-2016-best-practices-and-ethics-for-development-cooperation-projects/.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: A method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, 21:89–103.

Ellam Yua, Julie Raymond-Yakoubian, Raychelle Aluaq Daniel, and Carolina Behe. 2022. A framework for co-production of knowledge in the context of arctic research. *Ecology and Society*, 27(1). 34.

Isabelle Zaugg. 2019. Digital Surveillance and Digitally-disadvantaged Language Communities. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 227–230, Paris, France. European Language Resources Association (ELRA).

Ulrike Zeshan and Connie de Vos, editors. 2012. *Sign Languages in Village Communities: Anthropological and Linguistic Insights*. De Gruyter Mouton, Berlin, Boston.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. ChrEn: Cherokee-English Machine Translation for Endangered Language Revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595, Online. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2021. ChrEnTranslate: Cherokee-English Machine Translation Demo with Quality Estimation and Corrective Feedback. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 272–279, Online. Association for Computational Linguistics.

Appendix A

Data Statements Version 2 Schema and Best Practices

This appendix presents the data statements Version 2 schema and best practices. The schema, a guide for writing data statements following the Version 2 schema (Bender et al., 2021a), and data statements Version 2 templates are also available at `https://techpolicylab.uw.edu/data-statements/`.

A.1 General Best Practices

1. Remember that a broad range of people may be consulting data statements including but not limited to researchers within natural language processing, researchers in other fields (e.g., linguistics, law, or digital humanities), regulators, procurers, and members of and advocates for affected communities.
2. For datasets containing sensitive or proprietary information, whenever possible write the data statement so that it can be made publicly accessible (e.g., avoid including non-anonymized sensitive information).
3. Consider using the data statement elements as a checklist for dataset design.
4. Some of the data statement elements concern information that may require advanced planning to collect (e.g., demographic information). We recommend determining what information is to be collected and how at the start of the project, leaving time for ethics review board approval as appropriate.

5. For crafting your data statement, we recommend using an interview format with an external partner (e.g., someone not involved in the project). This is both fun and instructive. In effect, the external partner treats each data statement element as a question to be posed to a project member. In engaging with someone not involved in the construction of the dataset to discuss and clarify answers, you can get a good sense of what information and how much detail is needed in the data statement.

6. When using technical terms, make use of 15 Glossary.

7. When information is not known or unavailable, state this explicitly. It is valuable for readers to know, for example, that demographic information or information about specific language varieties is unavailable. Missing information is not a reason to forgo creating a data statement; clearly indicate what is missing and provide what information you can.

8. For datasets with extensive documentation outside the data statement (e.g., annotation guides), provide short summaries with pointers to the longer documents. It should be possible to know which key questions are answered in the other document(s).

9. Writing clear, concise data statements takes time and thought. We recommend iterating on the text of the data statement.

10. If the content of the dataset contains materials that could be a trigger for trauma, we recommend making a note of this in either 3 Curation Rationale or 14 Other.

11. If you reference papers and resources (aside from the dataset citation provided in 1 Header), include a reference list at the end of the data statement with full citations.

12. Once drafted, review your data statement for words or phrases used to describe speakers or their language varieties that might be experienced as diminishing and make revisions as appropriate.

13. Consider accessibility. When possible, use state of the art tools to check for accessibility, for example, for blind and low-vision readers.

14. Publish the data statements in the language(s) of the dataset, in addition to any languages of broader communication (such as English).

15. Provide the data statement together with the dataset. This is the canonical location for the most up to date version of the data statement. A link to the data statement along with 2 Executive Summary should be included in (1) any paper discussing the dataset or its uses and (2) the documentation for any system trained on the dataset. In publications presenting datasets, we recommend including the data statement as an appendix along with a pointer to where updated versions of the data statement may be found.

16. For datasets that are not publicly available (e.g., those containing non-anonymized health information or proprietary data), whenever possible make the data statement publicly accessible. See also General Best Practice 2 above.

A.2 Key Terms

Annotator refers to someone who assigns annotations to the raw language data, including transcribers of spoken or signed data.

Disordered speech refers to speech that has been affected by physiological conditions that affect a person's ability to produce speech sounds.

Elicited data refers to text that speakers were prompted to produce specifically for the purposes of constructing the dataset.

Found data refers to text that was produced by speakers for their own communicative purposes and collected after the fact for a dataset.

Language data refers to spoken, written or signed utterances.

Language variety refers to a manifestation of a given language (e.g., dialect); it does so without privileging one manifestation of the language as primary over others.

Speaker refers to someone who is competent in at least one modality for a language, meaning they are able to speak, sign and/or write in the language as well as perceive and understand speech, sign or text in it.

Speech refers to linguistic activity (i.e., the production of spoken, signed or written language).

Synthetic text refers to text produced by an algorithm rather than a person.

Text refers to a sequence of language data.

A.3 Schema Elements

1 HEADER

Why For dataset creators and data statement authors, this information ensures that credit and responsibility for the various documents are allocated appropriately.

For data statement readers, this information clarifies the source and authorship for the various documents pertaining to a dataset. Such information is particularly important when the author and source of the data statement differs from the author and source of the dataset, or when different versions of the data statement have different authors and sources.

What The header should include the following:

- Dataset Title
- Dataset Curator(s) [name, affiliation]
- Dataset Version [version, date]
- Dataset Citation and, if available, DOI
- Data Statement Author(s) [name, affiliation]
- Data Statement Version [version, date]
- Data Statement Citation
- Links to versions of this data statement in other languages

Best Practices

1. In order to manage updates over time, both datasets and their associated data statements should be versioned. That is, each updated dataset version should have its own updated data statement version. The data statement version number should be included in the data statement citation and is requested above. (Note that "Data Statement Version" refers to the version of the data statement, not the version of the data statement schema that is being used.)

2. In creating a standard citation for your data statement, we recommend including the following information about the data statement: authors, date, title, version, institution, and URL or DOI. The following is an example data statement citation:

   Gonzalez-Dios, Itziar. (2021). *Data Statement for the Corpus of Basque Simplified Texts*. Version 2. University of the Basque Country (UPV/EHU). http://www.ixa.eus/node/13302

3. Consider web accessibility and the longevity of data statement location (e.g., university archives or ACM digital library).

2 EXECUTIVE SUMMARY

Why For dataset creators, the executive summary provides the project team with a concise description of the dataset that can serve as a guiding statement of purpose throughout the dataset development. It can also be used in documents relating to the project, such as grant proposals, dissertation prospectuses, emails to potential collaborators, and project reports. A summary drafted before the data collection will need to be updated to reflect the final version.

For data statement readers, the executive summary provides a concise description of the dataset that can be used to make an initial determination about the appropriateness of the dataset for a specific purpose. The executive summary along with a pointer to the full data statement should be included in any publication using the dataset for training, tuning, or testing a system, and, as appropriate, for certain kinds of system documentation.

What The executive summary is a short (60–100 word) summary of the data statement that at a minimum should include: (1) a one-sentence description of the curation rationale, (2) the language(s), and (3) an overview of relevant quantitative information such as the dataset size.

Best Practices

1. We recommend finalizing the executive summary after the other elements have been drafted as that will help to clarify what level of detail is appropriate for this executive summary and which details are best included in other elements.

2. We recommend limiting the executive summary to descriptive facts about the dataset in and of itself (e.g., do not make comparisons to or assume familiarity with other datasets). Doing so will enable reuse over longer time periods (e.g., 20+ years).

Why For dataset creators, a curation rationale can help to promote intentionality in data selection and ensure representativeness. In addition, as difficult decisions arise, an explicit rationale can help to structure and resolve discussions about the data collection process and select pathways going forward.

For data statement readers, an explicit statement of why and how the dataset was curated can help with inferences about the domain of generalizability of systems trained on the dataset. Knowing which texts were included, and what the goals were in selecting texts, can be especially important in datasets too large to thoroughly inspect by hand.

What The curation rationale should answer questions including: Why was this dataset created? What is the task or research question the dataset is intended to address? Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection? What is the internal organization of the dataset? What constitutes a data instance?

Best Practices

1. If the dataset includes different categories of data (e.g., radio news and talk shows), include additional qualitative information describing the rationale for including different categories and their distribution within the larger dataset. Further data statement elements below should speak to each subcategory.

2. If the dataset involves subselection from a larger collection, specify topics, keywords, or other filters used and the reasons for choosing each. Technical details can be provided in 9 Preprocessing and Data Formatting.

3. We recommend writing the curation rationale after the other elements have been drafted. This will help to clarify what level of detail is appropriate for the curation rationale as well as which details are best included in other elements, thereby reducing repetition.

Why For dataset creators, the source dataset documentation can provide examples and language to draw from or reference when drafting the current data statement.

For data statement readers, the source dataset documentation can help with understanding how the current dataset builds upon and differs from the original task and data collection. Links to the source dataset show the user where to go look for further information, especially for the curation rationale of the source dataset.

What For datasets built out of pre-existing datasets, a link to a data statement for each source dataset should be included. If a data statement is not available, provide a link to a publication or other documentation. Provide links to licenses for source datasets, where applicable.

Best Practices

1. Include only immediate sources. For the situation where a chain of datasets have been built (e.g., A was the original source data set; B was built from A; C was built from B), then the data statement for the most current dataset (e.g., C) should only refer to the immediate source (e.g., B).

2. Include enough detail in the body of the data statement so that should the links between the data statement and the immediate source break, the data statement could function reasonably well as a stand-alone document.

Why Natural language processing algorithms embed assumptions about language structure; when applying an algorithm to a dataset from a language variety that differs structurally from that embedded in the algorithm unexpected behaviors may occur.

For dataset creators, a clear conception of the targeted language varieties can help inform decisions about data sources, curation, and annotation.

For data statement readers, accurate descriptions of the language varieties in the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third party technology developers or adopters to make similar assessments of match to populations at a future time.

What All of the languages and language varieties represented in the dataset should be characterized with (1) a language tag from BCP-47[1] identifying the language variety (e.g., en-US or yue-Hant-HK), and (2) a prose description elucidating and elaborating on the BCP-47 tag (e.g., English as spoken in Palo Alto, California; Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin).

Best Practices

1. Describe all language varieties represented in the dataset. For translation datasets, this would include both sides of the bitext. If the language variety used for annotations differs from the language variety of the source data, again document both.

2. Especially for less well studied languages, the description of the language variety should include enough information to situate it for dataset users unfamiliar with that variety. These descriptions should be written with respect and care to avoid harmful language ideologies (Kroskrity, 2005).

3. In the prose description, describe the dialects included in the dataset as accurately as possible with respect to national, regional and other sociolinguistic variation (e.g., rather than saying "American English", say "Standardized American English" or "Northeastern American English" as appropriate).

[1] https://tools.ietf.org/rfc/bcp/bcp47.txt

6 SPEAKER DEMOGRAPHIC

Why Beyond the language variety tied to a community of speakers (see 5 Language Varieties), individual speakers bring their own identities to their speech patterns. Specifically, sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with speaker demographic characteristics (Labov, 1966), as speakers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). In addition, when individuals speak a second language, properties of their first language affect their speech production in their second language (Ellis, 1994, Ch. 8). A further source of variation can be found in physiological sources such as disordered speech (e.g., dysarthria) (Christensen et al., 2012; Nicolao et al., 2016).

For dataset creators, a clear conception of the demographic categories targeted during the data collection process can help inform decisions about data sources, curation, and annotation. Data statements also enable the discovery of underserved populations across the overall data catalogue which, in turn, may influence choices for constructing the new dataset.

For data statement readers, accurate descriptions of the people represented in the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third party technology developers or adopters to make similar assessments of match to populations at a future time.

What All of the speaker groups represented in the dataset should be characterized with a prose description. Demographic categories are context- and culture-specific; therefore, locally appropriate categories and definitions should be used. Suggested specifications include:

- Age
- Gender
- Race/ethnicity
- Socioeconomic status
- First language(s)

- Proficiency in the language(s) of the data
- Number of different speakers represented
- Presence of disordered speech

Best Practices

1. Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).
2. Because the definitions and labels of demographic categories can change over time, include the dates when the data were produced and when the data were collected.
3. If the dataset includes speakers with different roles (e.g., interviewers, interviewees, and interpreters), provide demographic information for each role separately.
4. If the dataset consists entirely of synthetic text, if available, provide demographic information for the speakers in the training data for the automatic generation system.
5. If the dataset contains both found and elicited data, provide separate speaker demographics for each.
6. Be specific when describing demographic information, particularly with respect to category labels (e.g., rather than stating "all races" or "all ages" provide a set of labels or range of values) and source (e.g., self-reported vs. estimated).
7. When self-reported demographic data is not available, we recommend estimating demographic data by referring to studies of relevant larger populations (e.g., surveys of gender identities of Wikipedia editors) rather than trying to infer labels with classification tools (e.g., name-based gender attribution).
8. Report demographic information at the level of the entire dataset rather than attached to individual speakers to help protect their privacy.
9. When the number of participants and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect participant privacy.

7 ANNOTATOR DEMOGRAPHIC

Why Linguistic variation correlated with language users' demographics is also relevant for annotators. Specifically, annotators' own life experience influences their knowledge of language and how language is used by others and, thus, their perception of what they are annotating (Derczynski et al., 2016; Talat, 2016). As people annotate training datasets, they necessarily bring their perspectives to their annotations and, thereby, into the natural language processing models trained on that data.

For dataset creators, an accurate description of annotator demographics can be helpful in hiring annotators whose demographics closely match those of the speakers or, if that is not feasible, in identifying demographic gaps between annotators and speakers, and developing annotation guidelines accordingly, sensitive to those gaps.

For data statement readers, accurate descriptions of the annotators' demographics are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third party technology developers or adopters to make similar assessments of match to populations at a future time.

What All of the annotator groups represented in the dataset, including those who developed the guidelines, should be characterized with a prose description. Demographic categories are context- and culture-specific; therefore, locally appropriate categories and definitions should be used. Suggested specifications include:

- Age
- Gender
- Race/ethnicity
- Socioeconomic status
- First language(s)
- Proficiency in the language(s) of the data being annotated
- Number of different annotators represented
- Relevant training

Best Practices

1. Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).

2. Because the definitions and labels of demographic categories can change over time, include the dates when the annotations were produced.

3. If the dataset includes annotators with different roles (e.g., translators and labelers), provide demographic information for each role separately.

4. If the dataset includes automatically produced annotations, if available provide demographic information for the training data for the automatic annotation system.

5. If the dataset contains both found and elicited annotations, provide separate annotator demographics for each.

6. Be specific when describing demographic information, particularly with respect to category labels (e.g., rather than stating "all races" or "all ages" provide a set of labels or range of values) and source (e.g., self-reported vs. estimated).

7. When self-reported demographic data is not available, we recommend estimating demographic data by referring to studies of relevant larger populations (e.g., surveys of gender identities of Wikipedia editors) rather than trying to infer labels with classification tools (e.g., name-based gender attribution).

8. Report demographic information at the level of the entire dataset rather than attached to individual annotators to help protect their privacy.

9. When the number of annotators and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect annotator privacy.

8 SPEECH SITUATION AND TEXT CHARACTERISTICS

Why Characteristics of the speech situation can affect linguistic structure and patterns at many levels. For example, the intended audience of a linguistic performance can affect linguistic choices on the part of speakers. The time, place, and cultural context allow for deeper understanding of how the texts collected relate to their historical moment. Both genre and topic also influence the vocabulary and structural characteristics of texts (Biber, 1995).

For dataset creators, a clear conception of the targeted speech situation can help inform decisions about data sources, curation, and additional information to include through annotation (e.g., the timestamps of turn-taking in an asynchronous conversation).

For data statement readers, accurate descriptions of the speech situation in the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third party technology developers or adopters to make similar assessments of match to a target speech situation at a future time.

What A description of the speech situation in which the linguistic production occurred and/or the relevant text characteristics should be provided. This schema element may also be used to describe the cultural context of the language practices collected. Specifications include:

- Time and place of linguistic activity
- Date(s) of data collection
- Modality (spoken, signed, written)
- Scripted/edited vs. spontaneous
- Synchronous (e.g., in-person or live online chatting) vs. asynchronous (e.g., letters, emails, forums) interaction
- Speakers' intended audience
- Genre (e.g., newswire vs. social media)

- Topic (e.g., entertainment vs. natural disaster)
- Non-linguistic context (e.g., photos speakers were all looking at; a game participants are playing)
- Additional details about the cultural context (optional)

Best Practices

1. We recommend documenting as much of the speech situation and text characteristics information as possible before beginning the data collection. As the data is collected, update this information to reflect any changes.

Why For dataset creators, documenting the preprocessing procedure can help ensure that the procedure is applied consistently, especially when data is drawn from different sources or languages.

For data statement readers, this documentation can help clarify how changes introduced during preprocessing might affect system performance (e.g., replacing personal names with placeholders for anonymization, standardization of spelling, tokenization of sentences into words). Providing information about preprocessing also enables reproducible dataset construction.

What A description of all preprocessing and data formatting modifications made to the data (except for annotations) should be provided, including information about any anonymization procedures. The description should also specify which, if any, tools were used to make the modifications and whether the raw data is included in the dataset.

Best Practices

1. We recommend the description take the form of a list of ordered steps, with a link to external documentation of specific details, as appropriate. If different preprocessing steps are applied to different parts of the dataset, document each set of steps separately (e.g., adding whitespace only to scripts which do not usually use whitespace).

2. If the dataset is a filtered version of a larger data collection, we recommend using this schema element to provide technical detail on the specifics of the filters and their applications (e.g., specific search terms or filtering processes). This technical description of the filtering process complements the reasons for filtering provided in 3 Curation Rationale.

3. To the extent possible, provide software version information, citations, and links to repositories for the tools used in automatic processing.

10 CAPTURE QUALITY

Why For dataset creators, documenting quality issues can help inform decisions about preprocessing.

For data statement readers, accurate descriptions of the capture quality are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case (e.g., a corpus of collected speech may have word level transcription, but may not include disfluencies or mistakes made in the speech); and second, to enable future third party technology developers or adopters to make similar assessments of match to quality needs at a future time.

What A description of quality issues in data capture should be provided. This includes all types of quality issues that arise across a broad range of collection methodologies for capturing an otherwise impermanent event.

Best Practices

1. For data that include audiovisual recordings, describe the quality of the recording equipment and any aspects of the recording situation that could impact recording.
2. As appropriate, use this element to address other data quality concerns (e.g., image-to-text processing, granularity of transcription, or API reliability).

Why For dataset creators, it can be helpful to enumerate issues that have arisen for similar tasks or datasets as well as factors that might hinder the collection of a fully representative dataset. Ideally, this should be done before collecting data, in order to identify mitigation strategies. When setbacks occur in the course of creating a dataset, updating this schema element can help identify practical impacts on the resulting dataset and the extent to which the dataset in its current form meets its stated goal; such assessment can be helpful in guiding further data collection as appropriate.

For data statement readers, accurate descriptions of the challenges encountered in creating the dataset are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case; and second, to enable future third party technology developers or adopters to make similar assessments of match to populations at a future time.

What For any challenges that could not be fully addressed, a description of those challenges and characterization of the resulting limitations of the dataset should be provided.

Best Practices

1. We recommend documenting the challenges you encounter in the dataset development as they occur, including both the challenge and your strategy for addressing it.

2. For identifying possible limitations, we recommend using toolkits, such as Envisioning Cards[2] and the Lifecourse Checklist,[3] which guide practitioners to consider different populations and what representation means, as well as broader impacts.

3. We recommend noting any further precautions you would like future users of the dataset to be alert to.

[2] https://www.envisioningcards.com/

[3] https://docs.google.com/document/d/1uODpC40TQbD3VKjaorzSXY9Qc-Z9PB2qBf4SrwNiyOw/edit

12 METADATA

Why For dataset creators, it is important to be aware of and collect relevant metadata.

For data statement readers, data statements may be the "front door" through which they access the dataset. As such, it is important that the data statement contains pointers to the other metadata.

What A collection of pointers to relevant metadata should be provided. Suggestions include:

- License: Link to the license/copyright permissions for use or modification of the dataset
- Annotation Guidelines: Link to the published or online guidelines that annotators used to annotate the data
- Annotation Process: Link to documentation providing metadata about the annotation process, including protections for annotator anonymity, how annotators were compensated, and which aspects of the annotation were produced automatically
- Dataset Quality Metrics: Metrics for inter-annotator agreement and/or other numerical scores of dataset quality
- Errata: Link to the list of known errors and how to report additional ones

Best Practices

1. Include the most durable citations or links available (e.g., ISBN or DOI).
2. Include a link to the licensing/copyright permissions for both the dataset itself and the data curated to create the dataset.

Why For dataset creators, a clear conception of the terms of ethical approval can help inform decisions about data sources, curation, and annotation. Awareness of potential conflicts of interest can be helpful with managing or mitigating these.

For data statement readers, information about funding sources (which may have shaped curation and other decisions at the time of dataset creation) and ethical review (including the conditions of consent) may impact dataset selection.

What For projects supported by funding, a description of the funding source for the dataset and relevant information (e.g., grant number) should be specified. For projects that went through an ethical approval process, a link to the institution (e.g., IRB) should be provided. In addition, include: a brief description of any consent process used; if speakers or annotators were compensated, how compensation rates were determined; any access restrictions to the data; and any potential conflicts of interest.

Best Practices

1. If your data collection process involves a consent procedure, describe this element briefly with phrases such as “written consent”, “oral consent”, or “implied consent”.

2. If your institution does not have or require an ethical review process, we recommend stating this. Consider using a phrase such as “An institutional ethics review process was not accessible at the time of dataset creation.”

Why The data statement schema was designed to be broadly applicable to datasets containing language data, however there may be specific situations in which it would be useful to document other aspects of the dataset not covered by the schema.

What Any further considerations that are relevant for the dataset should be included here.

Best Practices

1. Avoid blurring the content boundaries of the established schema elements. If you identify a piece of information that does not fit in any of the other schema elements, include it here.

Why For data statement authors, using technical terms can make it easier to write efficient and precise documentation. Providing definitions for these technical terms can make the data statement accessible to a wider variety of audiences.

For data statement readers, definitions of technical terms can be especially important for three purposes: (1) understanding the intended use and limitations of the dataset, (2) conducting diagnostic analyses of system breakdowns, and (3) supporting the ability of impacted individuals, communities and their representatives to seek accountability for potential harms resulting from systems employing the dataset.

What A list of terms and associated definitions that may be technical or unfamiliar to non-experts should be provided.

Best Practices

1. We recommend engaging with someone outside of the project development team in order to determine what terms to include.

Appendix B

Collaborative Discussions for the Documentation and Design of Linguistic Archival Resources (C3DAR) Toolkit

The C3DAR toolkit is designed to support collaborative dataset curation and documentation between language communities and technical communities. It consists of general best practices, a list of key terms, and 17 schema elements corresponding to key considerations for designing datasets and writing documentation. Each schema element includes the rationale for its inclusion in the schema, its definition, and suggested best practices. By filling out each of the schema elements with a future dataset in mind, the dataset design team can thoroughly discuss plans for the dataset's content, creation process, and publication while also producing an initial draft of the dataset documentation. This process is intended to be iterative, with schema elements being drafted as decisions are made and updated as the project develops. The C3DAR toolkit and template is also available for download at `https://digital.lib.washington.edu/researchworks/handle/1773/50585`.

B.1 General Best Practices for Collaboration

1. Collaboration requires honesty, respect and care for team members, the community, and the community's history and values (South African San Institute, 2017; Hudson et al., 2010). Be mindful of asymmetrical power relations throughout the project within this historical context and how these interact with community cultural norms (Harris et al., 2009; Ontario Federation of Indigenous Friendship Centres, 2016).

2. Whenever possible, communicate in the language the community prefers. Relying on interpretation services may affect the results of the project and the research team's ability to understand the community's perspective, so it is best if at least one team member is able to communicate in the language of the community.

3. The project should center the needs and understandings of the community. The community should be involved in determining the project goals, methods, and evaluation criteria. The community's knowledge and ways of knowing are valid without reaffirmation via mainstream understandings and analysis (Ontario Federation of Indigenous Friendship Centres, 2016).

4. Be aware of the relevant axes of diversity within the community. Community representatives should reflect the community diversity, which may be uniquely defined depending on what demographic information and personal characteristics are most salient to the community. Be transparent in any recruiting processes. Recruitment of particular community members should be transparent as to why those community members were selected for the role so as not to create mistrust from the community with respect to the project or negatively impact the recruited community member (World Federation of the Deaf Expert Group on Developing Countries, 2016).

5. Allow time for negotiation processes according to community customs as well as for feedback and reviewing processes. While community collaborators may be aware of this difference, communicating about expected time frames of both academic and community processes can help the project members to prepare ahead of time for setbacks or find other ways to make use of time spent waiting (Coeur d'Alene Tribe of Idaho and University of Idaho, 2015).

6. Discuss the benefits that all parties will derive from the dataset, related projects, and the collaboration itself. In particular, the outcomes should include tangible and meaningful benefits to the community that address their self-identified needs. Consider whether commercialization will be allowed on the dataset or products derived from the dataset, and if so, how the benefits and responsibilities of commercialization will be managed (Argumedo et al., 2011).

7. The community and its knowledge should be protected against risks related to the project or resulting from later use of the dataset. The community members may need to be protected from physical and psychological harm, disparagement or disrespect, and confidentiality breaches. Community knowledge may be deemed sensitive and therefore inappropriate to include in any publications or publicly distributed data. Discuss the possible risks and develop mitigation strategies with the community. The implication is not that the collaboration team should be able to foresee all harms, but rather that active measures should be put in place to prevent harms and assess risks.

8. The community should be involved in developing culturally appropriate procedures for ongoing free, prior, informed and educated consent. This may include the language(s) that the procedure will be available in, whether a written version will be available, and how to make the project methods, potential risks and benefits, and confidentiality procedures clear to the community. Avoid assuming that potential benefits are obvious, making exaggerated claims, and understating the potential risks. The community should decide whether this consent is individual or collective.

9. Ownership of the community's knowledge, cultural heritage and data belongs with the community. Copies of the dataset and any other products should be returned to the community physically and/or in an accessible format. Discuss the ownership and management of the project deliverables and document the terms in schema element 14 Distribution.

10. Plan to meet periodically with collaborators to discuss updates and relevant questions. Establish a mediation process for handling disagreements as they arise.

11. Share updates with the community in a way that is transparent and comprehensible to those outside the project.

12. Each member of the project team should receive acknowledgement and due credit for their contributions to the project in a way that is meaningful to the team member.

B.2 General Best Practices for Documentation

1. Remember that a broad range of people may be consulting this documentation including but not limited to researchers within natural language processing, researchers in other fields (e.g., linguistics, law, or digital humanities), regulators, procurers, and members of and advocates for affected communities.

2. For datasets that will contain sensitive or proprietary information, whenever possible write the documentation so that it can be made publicly accessible (e.g., avoid including non-anonymized sensitive information).

3. Some of the elements concern information that may require advanced planning to collect (e.g., demographic information). We recommend determining what information is to be collected and how at the start of the project, leaving time for ethics review board approval as appropriate. Consult with communities early about appropriate demographic categories.

4. For refining your documentation, we recommend using an interview format with an external partner (e.g., someone not involved in the project). This is both fun and instructive. In effect, the external partner treats each element as a question to be posed to a project member. In engaging with someone not involved in the construction of the dataset to discuss and clarify answers, you can get a good sense of what information and how much detail is needed in the documentation.

5. When using technical terms, make use of 17 Glossary.

6. When information is not known or unavailable, state this explicitly. It is valuable for readers to know, for example, that demographic information or information about specific language varieties is unavailable. Missing information is not a reason to forgo creating documentation; clearly indicate what is missing and provide what information you can.

7. For datasets with extensive documentation outside this document (e.g., annotation guides), provide short summaries with pointers to the longer documents. It should be possible to know which key questions are answered in the other document(s).

8. Writing clear, concise documentation takes time and thought. We recommend iterating on the text of the documentation development.

9. If the content of the dataset will contain materials that could be a trigger for trauma, we recommend making a note of this in either 3 Curation Rationale or 16 Other.

10. If you reference papers and resources (aside from the dataset citation provided in 1 Header), include a reference list at the end of the documentation with full citations.

11. Once drafted, review your documentation for words or phrases used to describe language users or their language varieties that might be experienced as diminishing and make revisions as appropriate.

12. Consider accessibility. When possible, use state of the art tools to check for accessibility, for example, for blind and low-vision readers.

13. For datasets concerning languages other than English, also publish the documentation in the language(s) of the dataset.

14. Provide the documentation together with the dataset. This is the canonical location for the most up to date version of the documentation. 2 Executive Summary along with a link to the documentation should be included in (1) any paper discussing the dataset or its uses and (2) the documentation for any system trained on the dataset. In publications presenting datasets, we recommend including the documentation as an appendix along with a pointer to where updated versions of the documentation may be found.

15. For datasets that will not be publicly available (e.g., those containing non-anonymized health information or proprietary data), whenever possible make the documentation publicly accessible. See also General Best Practice for Documentation 2 above.

B.3 Key Terms

Annotator refers to someone who assigns annotations to the raw language data, including transcribers of spoken or signed data.

Disordered speech or sign refers to speech or sign that has been affected by physiological conditions that affect a person's ability to produce speech sounds or signs.

Elicited data refers to text that language users were prompted to produce specifically for the purposes of constructing the dataset.

Found data refers to text that was produced by language users for their own communicative purposes and collected after the fact for a dataset.

Language data refers to spoken, written or signed utterances.

Language user refers to someone who is competent in at least one modality for a language, meaning they are able to speak, sign and/or write in the language as well as perceive and understand speech, sign or text in it.

Language variety refers to a manifestation of a given language (e.g., dialect); it does so without privileging one manifestation of the language as primary over others.

Synthetic text refers to text produced by an algorithm rather than a person.

Text refers to a sequence of language data.

B.4 Schema Elements

1 HEADER

Why For dataset creators and documentation authors, this information ensures that credit and responsibility for the various documents are allocated appropriately.

For documentation readers, this information clarifies the source, authorship, and contributions for the various documents pertaining to a dataset. Such information is particularly important when the source, authors, and contributors of the documentation differs from the source, authors, and contributors of the dataset, or when different versions of the documentation have different sources, authors, and contributors.

What The header should include the following:

- Dataset Title
- Dataset Contributor(s) [name, affiliation, role]
- Dataset Version [version, date]
- Dataset Citation and DOI
- Documentation Contributor(s) [name, affiliation, role]
- Documentation Version [version, date]
- Documentation Citation
- Links to versions of this documentation in other languages

Best Practices

1. In order to manage updates over time, both datasets and their associated documentation should be versioned. That is, each updated dataset version should have its own updated documentation version. The documentation version number should be included in the documentation citation and is requested above. (Note that "Documentation Version" refers to the version of the documentation, not the version of the documentation schema that is being used.)

2. In creating a standard citation for your documentation, we recommend including the following information about the documentation: authors, date, title, version, institution, and URL or DOI.

3. Consider web accessibility and the longevity of documentation location (e.g., university archives or a community-owned repository). See 15 Maintenance for further considerations.

4. Discuss with community partners how they would prefer to be acknowledged for their contributions. For some communities, coauthorship is appropriate, while others may have another preferred method. Consider also how to acknowledge contributions such as consultations on local knowledge, reviewing materials, and other efforts supporting the development of the project.

2 EXECUTIVE SUMMARY

Why For dataset creators, the executive summary provides the project team with a concise description of the dataset that can serve as a guiding statement of purpose throughout the dataset development. It can also be used in documents relating to the project, such as grant proposals, dissertation prospectuses, emails to potential collaborators, and project reports to the community. A summary drafted before the data collection will need to be updated to reflect the final version.

For documentation readers, the executive summary provides a concise description of the dataset that can be used to make an initial determination about the appropriateness of the dataset for a specific purpose. The executive summary along with a pointer to the full documentation should be included in any publication using the dataset for training, tuning, or testing a system, and, as appropriate, for certain kinds of system documentation.

What The executive summary is a short (60–100 word) summary of the documentation that at a minimum should include: (1) a one-sentence description of the curation rationale, (2) the language(s), (3) an overview of relevant quantitative information such as the anticipated dataset size, and (4) a short description of how the community has been involved in the project.

Best Practices

1. We recommend finalizing the executive summary after the other elements have been drafted as that will help to clarify what level of detail is appropriate for this executive summary and which details are best included in other elements.

2. We recommend limiting the executive summary to descriptive facts about the dataset in and of itself (e.g., do not make comparisons to or assume familiarity with other datasets). Doing so will enable reuse over longer time periods (e.g., 20+ years).

Why For dataset creators, a curation rationale can help to promote intentionality in data selection and ensure representativeness. In addition, as difficult decisions arise, an explicit rationale can help to structure and resolve discussions about the data collection process and select pathways going forward.

For documentation readers, an explicit statement of why and how the dataset was curated can help with inferences about the domain of generalizability of systems trained on the dataset. Knowing which texts were included, and what the goals were in selecting texts, can be especially important in datasets too large to thoroughly inspect by hand.

What The curation rationale should answer questions including: What is the intended purpose of this dataset? What is the task or research question the dataset is intended to address? Which texts will be included and what are the goals in selecting texts, both in the original collection and in any further subselection? What will be the internal organization of the dataset? What will constitute a data instance? How will the dataset support community goals?

Best Practices

1. If the dataset will include different categories of data (e.g., radio news and talk shows), include additional qualitative information describing the rationale for including different categories and their distribution within the larger dataset. Further elements below should speak to each subcategory.

2. If the dataset will involve subselection from a larger collection, specify topics, keywords, or other filters that will be used and the reasons for choosing each. Technical details can be provided in 9 Preprocessing and Data Formatting.

3. We recommend finalizing the curation rationale after the other elements have been drafted. This will help to clarify what level of detail is appropriate for the curation rationale as well as which details are best included in other elements, thereby reducing repetition.

Why For dataset creators, the source dataset design and documentation can provide examples and language to draw from or reference when drafting the current dataset design and documentation.

For documentation readers, the source dataset documentation can help with understanding how the current dataset will build upon and differ from the original task and data collection. Links to the source dataset show the user where to go look for further information, especially for the curation rationale of the source dataset.

What For datasets that will be built out of pre-existing datasets, a link to the documentation for each source dataset should be included. Provide links to licenses, copyright, or terms of use for source datasets, where applicable.

Best Practices

1. Include only immediate sources. For the situation where a chain of datasets have been built (e.g., A was the original source data set; B was built from A; C was built from B), then the documentation for the most current dataset (e.g., C) should only refer to the immediate source (e.g., B).

2. Include enough detail in the body of the documentation so that should the links between the documentation and the immediate source break, the documentation could function reasonably well as a stand-alone document.

3. If the source dataset was collected under specific consent conditions, ensure that those conditions allow for further reuse and distribution as needed by the current dataset. When in doubt, contact the source dataset manager and ask about developing a new opt-in consent procedure for the language users who created the source data to agree to the new use and dissemination of their data.

Why Natural language processing algorithms embed assumptions about language structure; when applying an algorithm to a dataset from a language variety that differs structurally from that embedded in the algorithm, unexpected behaviors may occur.

For dataset creators, a clear conception of the targeted language varieties can help inform decisions about data sources, curation, and annotation.

For documentation readers, accurate descriptions of the language varieties in the dataset are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, to enable future third party technology developers or adopters to make similar assessments of match to a target linguistic situation at a future time.

What All of the languages and language varieties that will be represented in the dataset should be characterized with (1) a language tag from BCP-47[1] identifying the language variety (e.g., en-US or yue-Hant-HK), and (2) a prose description elucidating and elaborating on the BCP-47 tag (e.g., English as spoken in Palo Alto, California; Cantonese written with traditional characters by speakers in Hong Kong who are bilingual in Mandarin; French Sign Language as used in Marseille, France).

Best Practices

1. Describe all language varieties that will be represented in the dataset and the metadata. For translation datasets, this would include both sides of the bitext. If the language variety that will be used for annotations differs from the language variety of the source data, again document both.

2. Especially for less well studied languages, the description of the language variety should include enough information to situate it for dataset users unfamiliar with that variety. These descriptions should be written with respect and care for how the community would like their language to be known and avoid harmful language ideologies (Kroskrity, 2005).

[1] `https://tools.ietf.org/rfc/bcp/bcp47.txt`

3. In the prose description, describe the dialects that will be included in the dataset as accurately as possible with respect to national, regional and other sociolinguistic variation (e.g., rather than saying “American English”, say “Standardized American English” or “Northeastern American English” as appropriate).

6 LANGUAGE USER DEMOGRAPHIC

Why Beyond the language variety tied to a community of speakers or signers (see 5 Language Varieties), individual language users bring their own identities to their linguistic patterns. Specifically, sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with the language user's demographic characteristics (Labov, 1966; Kusters and Lucas, 2022), as speakers and signers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). In addition, when individuals speak or sign a second language, properties of their first language affect their production in their second language (Ellis, 1994, Ch. 8; Quinto-Pozos, 2008). A further source of variation can be found in physiological sources such as disordered speech or sign (e.g., dysarthria) (Christensen et al. 2012, Nicolao et al. 2016; for dysarthria in signed languages see Tyrone 2014).

For dataset creators, a clear conception of the demographic categories targeted during the data collection process can help inform decisions about data sources, curation, and annotation. Documentation can also enable the discovery of underserved populations across the overall data catalog which, in turn, may influence choices for constructing the new dataset.

For documentation readers, accurate descriptions of the people represented in the dataset are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, to enable future third party technology developers or adopters to make similar assessments of match to a target linguistic situation at a future time.

What All of the language user groups that will be represented in the dataset should be characterized with a prose description. Demographic categories are context- and culture-specific; therefore, locally appropriate categories and definitions should be used as determined by the community. Suggested specifications include:

- Age
- Gender
- Race/ethnicity

- Socioeconomic status
- First language(s)
- Proficiency in the language(s) of the data
- Proposed number of different speakers or signers represented
- Presence of disordered speech or sign

Best Practices

1. Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).
2. Because the definitions and labels of demographic categories can change over time, include the dates when the data were or will be produced and when the data will be collected.
3. If the dataset will include language users with different roles (e.g., interviewers, interviewees, and interpreters), provide demographic information for each role separately.
4. If the dataset will consist entirely of synthetic text, if available, provide demographic information for the language users in the training data for the automatic generation system.
5. If the dataset will contain both found and elicited data, provide separate language user demographics for each.
6. Be specific when describing demographic information, particularly with respect to category labels (e.g., rather than stating "all races" or "all ages" provide a set of labels or range of values) and source (e.g., self-reported vs. estimated).
7. When self-reported demographic data is not available, we recommend estimating demographic data by referring to studies of relevant larger populations (e.g., surveys of gender identities of Wikipedia editors) rather than trying to infer labels with classification tools (e.g., name-based gender attribution).
8. Report demographic information at the level of the entire dataset rather than attached to individual language users to help protect their privacy.

9. When the number of participants and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect participant privacy. Discuss with community representatives what demographic information may be safely gathered and shared.

7 ANNOTATOR DEMOGRAPHIC

Why Linguistic variation correlated with the language user's demographics is also relevant for annotators. Specifically, the annotators' own life experience influences their knowledge of language and how language is used by others and, thus, their perception of what they are annotating (Derczynski et al., 2016; Talat, 2016). As people annotate training datasets, they necessarily bring their perspectives to their annotations and, thereby, into the natural language processing models trained on that data.

For dataset creators, an accurate description of annotator demographics can be helpful in hiring annotators whose demographics closely match those of the language users in the dataset or, if that is not feasible, in identifying demographic gaps between annotators and language users in the dataset, and developing annotation guidelines accordingly, sensitive to those gaps.

For documentation readers, accurate descriptions of the annotators' demographics are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, to enable future third party technology developers or adopters to make similar assessments of match to a target linguistic situation at a future time.

What All of the annotator groups that will be represented in the dataset, including those who will develop the guidelines, should be characterized with a prose description. Demographic categories are context- and culture-specific; therefore, locally appropriate categories and definitions should be used as determined by the community. Suggested specifications include:

- Age
- Gender
- Race/ethnicity
- Socioeconomic status
- First language(s)
- Proficiency in the language(s) of the data being annotated

- Proposed number of different annotators represented
- Relevant training

Best Practices

1. Discussions of demographic categories should be informed by current best practice (e.g., as of 2021, for gender see Larson 2017).
2. Because the definitions and labels of demographic categories can change over time, include the dates for when the annotations will be produced.
3. If the dataset will include annotators with different roles (e.g., translators and labelers), provide demographic information for each role separately.
4. If the dataset will include automatically produced annotations, if available provide demographic information for the training data for the automatic annotation system.
5. If the dataset will contain both found and elicited annotations, provide separate annotator demographics for each.
6. Be specific when describing demographic information, particularly with respect to category labels (e.g., rather than stating "all races" or "all ages" provide a set of labels or range of values) and source (e.g., self-reported vs. estimated).
7. When self-reported demographic data is not available, we recommend estimating demographic data by referring to studies of relevant larger populations (e.g., surveys of gender identities of Wikipedia editors) rather than trying to infer labels with classification tools (e.g., name-based gender attribution).
8. Report demographic information at the level of the entire dataset rather than attached to individual annotators to help protect their privacy.
9. When the number of annotators and/or the community being sampled is small, we recommend reporting demographic information as a range to help protect annotator privacy. Discuss with community representatives what demographic information may be safely gathered and shared.

Why Characteristics of the linguistic situation can affect linguistic structure and patterns at many levels. For example, the intended audience of a linguistic performance can affect linguistic choices on the part of speakers, signers, and authors. The time, place, and cultural context allow for deeper understanding of how the language data collected relate to their historical moment. Both genre and topic also influence the vocabulary and structural characteristics of language data (Biber, 1995).

For dataset creators, a clear conception of the targeted linguistic situation can help inform decisions about data sources, curation, and additional information to include through annotation (e.g., the timestamps of turn-taking in an asynchronous conversation).

For documentation readers, accurate descriptions of the linguistic situation in the dataset are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case; and third, to enable future third party technology developers or adopters to make similar assessments of match to a target linguistic situation at a future time.

What A description of the situation in which the linguistic production will occur and/or the relevant text characteristics should be provided. This schema element may also be used to describe the cultural context of the language practices that will be collected. Specifications include:

- Time and place of linguistic activity
- Proposed date(s) of data collection
- Modality (spoken, signed, written)
- Scripted/edited vs. spontaneous
- Synchronous (e.g., in-person or live online chatting) vs. asynchronous (e.g., letters, emails, forums) interaction
- Language users' intended audience

- Genre (e.g., newswire vs. social media)
- Topic (e.g., entertainment vs. natural disaster)
- Non-linguistic context (e.g., photos participants were all looking at; a game participants are playing)
- Additional details about the cultural context (optional)

Best Practices

1. We recommend documenting as much of the linguistic situation and text characteristics information as possible before beginning the data collection. As the data is collected, update this information to reflect any changes.
2. When describing the cultural context, use community vocabulary, concepts, and interpretations to convey the cultural significance, when deemed appropriate for public dissemination by the community.

Why For dataset creators, documenting the preprocessing procedure can help ensure that the procedure is applied consistently, especially when data is drawn from different sources or languages.

For documentation readers, this documentation can help clarify how changes introduced during preprocessing might affect system performance (e.g., replacing personal names with placeholders for anonymization, standardization of spelling, tokenization of sentences into words). Providing information about preprocessing also enables reproducible dataset construction.

What A description of all preprocessing and data formatting modifications that will be made to the data (except for annotations) should be provided, including information about any anonymization procedures. The description should also specify which, if any, tools will be used to make the modifications and whether the raw data will be included in the dataset.

Best Practices

1. We recommend the description take the form of a list of ordered steps, with a link to external documentation of specific details, as appropriate.

2. If different preprocessing steps will be applied to different parts of the dataset, document each set of steps separately (e.g., adding whitespace only to scripts which do not usually use whitespace).

3. If the dataset will be a filtered version of a larger data collection, we recommend using this schema element to provide technical detail on the specifics of the filters and their applications (e.g., specific search terms or filtering processes). This technical description of the filtering process complements the reasons for filtering provided in 3 Curation Rationale.

4. To the extent possible, provide software version information, citations, and links to repositories for the tools that will be used in automatic processing.

5. When anonymizing video or image data, modifications to the data such as blurring faces may remove necessary linguistics context and information, especially for signed languages. If language users in the

dataset have not agreed to public dissemination of their video or image data without anonymization, consider all available methods for protecting the language users' privacy, such as access restrictions, and ensuring the usefulness of the dataset for the community.

Why For dataset creators, documenting quality issues can help inform decisions about preprocessing.

For documentation readers, accurate descriptions of the recording quality are important for at least three reasons: first, to support community assessments, comparisons, and cataloging of currently available data; second, to assess if the dataset would be well-matched for a particular intended use case (e.g., a corpus of collected speech may have word level transcription, but may not include disfluencies or mistakes made in the speech); and third, to enable future third party technology developers or adopters to make similar assessments of match to quality needs at a future time.

What A description of anticipated quality issues in data capture should be provided. This includes all types of quality issues that arise across a broad range of collection methodologies for capturing an otherwise impermanent event.

Best Practices

1. For data that will include audiovisual recordings, describe the quality of the recording equipment and any aspects of the recording situation that could impact recording.

2. As appropriate, use this element to address other data quality concerns (e.g., image-to-text processing, granularity of transcription, or API reliability).

Why For dataset creators, it can be helpful to enumerate issues that have arisen for similar tasks or datasets as well as factors that might hinder the collection of a fully representative dataset. Ideally, this should be done before collecting data, in order to identify mitigation strategies. When setbacks occur in the course of creating a dataset, updating this schema element can help identify practical impacts on the resulting dataset and the extent to which the dataset in its current form meets its stated goal; such assessment can be helpful in guiding further data collection as appropriate.

For documentation readers, accurate descriptions of the challenges encountered in creating the dataset are important for at least two three reasons: first, to evaluate the degree to which this dataset has contributed towards community goals; second, to assess if the dataset would be well-matched for a particular intended use case; and third, to enable future third party technology developers or adopters to make similar assessments of match to populations at a future time.

What For any anticipated challenges that may not be fully addressed, a description of those challenges and characterization of the potential resulting limitations of the dataset should be provided.

Best Practices

1. We recommend documenting the challenges you encounter in the dataset development as they occur, including both the challenge and your strategy for addressing it.

2. For identifying possible limitations, we recommend using toolkits, such as Envisioning Cards[2] and the Lifecourse Checklist,[3] which guide practitioners to consider different populations and what representation means, as well as broader impacts.

3. We recommend noting any further precautions you would like future users of the dataset to be alert to.

[2] https://www.envisioningcards.com/

[3] https://docs.google.com/document/d/1uODpC40TQbD3VKjaorzSXY9Qc-Z9PB2qBf4SrwNiyOw/edit

Why For dataset creators, it is important to be aware of and collect relevant metadata.

For documentation readers, documentation may be the "front door" through which they access the dataset. As such, it is important that the documentation contains pointers to the other metadata.

What A collection of pointers to relevant metadata should be provided. Suggestions include:

- Annotation Guidelines: Link to the published or online guidelines that annotators will use to annotate the data
- Annotation Process: Link to documentation providing metadata about the proposed annotation process, including protections for annotator anonymity, how annotators will be compensated, and which aspects of the annotation will be produced automatically
- Dataset Quality Metrics: Proposed metrics for inter-annotator agreement and/or other numerical scores of dataset quality

Best Practices

1. Include the most durable citations or links available (e.g., ISBN or DOI).

Why For dataset creators, a clear conception of the terms of the ethical approval can help inform decisions about data sources, curation, and annotation. Awareness of potential conflicts of interest can be helpful with managing or mitigating these. If a community has an ethical review process, engagement with this process can help surface community-specific concerns with the dataset creation and help guide the dataset creation to support community goals.

For documentation readers, information about funding sources (which may shape curation and other decisions at the time of dataset creation) and ethical review (including the conditions of consent) may impact dataset selection.

What For projects supported by funding, a description of the funding source for the dataset and relevant information (e.g., grant number) should be specified. For projects that went through an ethical approval process, a link to the institution (e.g., IRB) should be provided. In addition, include: a brief description of any proposed consent process; if language users in the dataset or annotators will be compensated, how compensation rates will be determined; and any potential conflicts of interest.

Best Practices

1. If your data collection process will involve a consent procedure, describe this element briefly with phrases such as “written consent”, “oral consent”, or “implied consent”.

2. If your institution does not have or require an ethical review process, we recommend stating this. Consider using a phrase such as “An institutional ethics review process will not be accessible at the time of dataset creation.”

3. If the community has an ethical review process, we recommend stating whether or not the project has engaged with the process and any results from the engagement.

Why For dataset creators, having a detailed plan for distribution can help inform data curation decisions as it determines whether the team should only collect data that will allow for public distribution or if access to the dataset or parts of the dataset will be restricted. The data collection team will also need to provide distribution information to the people the data is collected from as part of the consent procedure.

For documentation readers, a detailed description of the permitted uses of this dataset can help in determining whether the dataset is suitable for a particular use case and whether the dataset can be further redistributed. If the documentation is the first access point for a reader, the distribution explanation can help the reader find and access the dataset or explain why they are unable to find or access the dataset. Documentation that communicates planned revisions or removal of the dataset in advance may also help documentation readers prepare for changes to the dataset.

What A description of how the dataset will be distributed should be specified. This includes the method of distribution (e.g., through a data archive, files on website, API, GitHub) and any access restrictions on the dataset or subsets of the dataset (e.g. sensitive or confidential content, intellectual property (IP)-based restrictions, export controls, or other regulatory restrictions). If the dataset or portions of the dataset will be distributed under an IP license, copyright, or terms of use (ToU), describe the licenses, copyright, and/or ToU. Provide links or other access points to, or otherwise reproduce, any relevant licensing terms or ToU, and list any fees associated with these restrictions. Other suggestions for detailing the distribution plan include:

- Who the dataset will be distributed to (e.g. third parties outside of the entity (community, company, institution, or organization) on behalf of which the dataset was created)
- If there are conditions for accessing the dataset or subsets of the dataset, and if so, what the conditions for being granted access are
- If the dataset will have a digital object identifier (DOI)
- When the dataset will be distributed

Best Practices

1. Review the data with community representatives to determine which portions of the dataset will be culturally appropriate to share broadly, which portions should be restricted to relevant groups, and which portions should be accessible to community members only.

2. In addition to the general distribution method, provide the community with a locally accessible copy of the dataset.

3. When choosing terms for a license, copyright, or ToU, consider uses that will be allowed as well as uses that will be disallowed. The community should decide on whether they want to allow third-party uses such as research, use in court, technical development, and commercialization.

Why For dataset creators, a maintenance plan for the dataset may help to ensure that the dataset will continue to be usable and accessible to both the community and other intended audiences. For communities, developing a maintenance plan may help in considering archiving options along with their benefits, risks, and costs prior to data collection.

For documentation readers, information about the dataset's maintenance will help determine who to contact for questions about the dataset after it has been published. Information about previous updates may help determine which version of the dataset will be most applicable to the reader's use case and help the reader plan for integrating dataset updates into their system development.

What A description of how the dataset will be maintained should be specified. This includes who will support, host, and maintain the dataset and what the proposed method for contacting the manager of the dataset will be. Other considerations include:

- If and where a list of errors found after the dataset's publication will be maintained and how to report errors
- How often, by whom, and how updates to the dataset (e.g., to correct labeling errors, add new data, delete data) will be communicated to users (e.g., mailing list, GitHub)
- Applicable limits on the retention of the data associated with the instances (e.g., will individuals in question be told that their data will be retained for a fixed period of time and then deleted) and how those limits will be enforced
- Whether older versions of the dataset will continue to be supported, hosted, and maintained
- How users will be notified that the dataset is outdated or no longer available
- Whether others will be able to extend/augment/build on/contribute to the dataset, and if so, how others will be able to contribute, if and how these contributions will be validated, and whether these contributions will be further communicated and distributed to other users

Best Practices

1. Consider web accessibility and the longevity of dataset location (e.g., university archives or a community-owned repository), especially with respect to how community members will access the data.

2. We recommend having a process for removing data, in the event that someone would like to have their data or community-sensitive data removed from the dataset.

Why This toolkit was designed to be broadly applicable to datasets containing language data, however there may be specific situations in which it would be useful to document other aspects of the dataset not covered by the schema.

What Any further considerations that are relevant for the dataset should be included here.

Best Practices

1. Avoid blurring the content boundaries of the established schema elements. If you identify a piece of information that does not fit in any of the other schema elements, include it here.

Why For documentation authors, using technical terms can make it easier to write efficient and precise documentation. Using local terminology throughout the documentation centers the community's understanding of the data and its cultural significance. Providing definitions for these technical terms can make the data statement accessible to a wider variety of audiences.

For documentation readers, definitions of technical terms can be especially important for three purposes: (1) understanding the intended use and limitations of the dataset, (2) conducting diagnostic analyses of system breakdowns, and (3) supporting the ability of impacted individuals, communities and their representatives to seek accountability for potential harms resulting from systems employing the dataset. Definitions of local vocabulary can be important for understanding and interpreting the data in community-appropriate ways and acknowledging the validity of community knowledge and ways of knowing.

What A list of terms and associated definitions that may be technical or unfamiliar to non-experts should be provided.

Best Practices

1. We recommend engaging with someone outside of the project development team in order to determine what terms to include.

Appendix C

Coding Manual for the Retrospective Investigation

The following is the coding manual I used to label the guidelines and licenses in the retrospective investigation. Starting from the International Society of Ethnobiology (ISE) Code of Ethics (International Society of Ethnobiology, 2006 with 2008 additions) used by Tunón et al. (2016) to code their set of ethical guidelines, I edited the principles to apply to more general research and to address communities as well as researchers. See Section 5.3.1 for details on the changes made from the original ISE Code of Ethics.

Acknowledgement and Due Credit (A&DC) This principle recognises that Indigenous peoples, traditional societies and local communities must be acknowledged in accordance with their preference and given due credit in all agreed publications and other forms of dissemination for their tangible and intangible contributions to research activities. Co-authorship should be discussed with all contributors. Acknowledgement and due credit to Indigenous peoples, traditional societies and local communities extend equally to secondary or downstream uses and applications and researchers will act in good faith to ensure the connections to original sources of knowledge and resources are maintained in the public record.

Active Protection (AP) This principles recognises the importance of both community leaders, community researchers, and outside researchers taking active measures to protect Indigenous peoples, traditional societies and local communities traditional knowledge and their rights with respect to that knowledge. This

principle encourages research that enhances the relationships of Indigenous peoples, traditional societies and local communities with their traditional knowledge and thereby promotes the maintenance of cultural and biological diversity.

Collaboration and Active Participation (C&AP) This principle recognises the crucial importance of collaboration between Indigenous peoples, traditional societies and local communities and research partners. Community collaborators and leaders should actively contribute in all phases of research and related activities from inception to completion, as well as in application of research results. This includes collaboration on research design to address local needs and priorities, and prior review of results before publication or dissemination to ensure accuracy of information and adherence to the standards represented by this Code of Ethics.[1]

Collective and Individual Inalienability (C&II) This principle recognises the inalienable rights of Indigenous peoples, traditional societies and local communities in relation to their traditional territories and the natural resources (including biological and genetic resources) within them and associated traditional knowledge. These rights are collective by nature but can include individual rights. It shall be for Indigenous peoples, traditional societies and local communities to determine for themselves the nature, scope and alienability of their respective resource rights regimes.

Confidentiality (C) This principle recognises that Indigenous peoples, traditional societies and local communities, at their sole discretion, have the right to exclude from publication and/or to have kept confidential any information concerning their culture, identity, language, traditions, mythologies, spiritual beliefs or genomics. Parties to the research have a responsibility to be aware of and comply with local systems for management of knowledge and local innovation, especially as related to sacred and secret knowledge. Furthermore, such confidentiality shall be guaranteed by researchers and other potential users. Indigenous peoples, traditional societies and local communities also have the rights to privacy and anonymity, at their discretion.

[1] "This Code of Ethics" is the International Society of Ethnobiology (2006 with 2008 additions) referring to itself.

Diligence (D) This principle recognises that researchers are expected to have a working understanding of the local context prior to entering into research relationships with a community. This understanding includes knowledge of and willingness to comply with local governance systems, cultural laws and protocols, social customs and etiquette. Researchers are expected to conduct research in the local language to the degree possible, which may involve language fluency or employment of interpreters

Diversity and Representation (D&R) This principle recognizes the diverse experiences, understandings, and way of life that reflect Indigenous peoples, traditional societies and local communities' contemporary cultures. This principle also recognizes that community and culture is dynamic and acknowledges the changes that may occur within a culture over time. As a result of this change, local communities are likely to hold some degree of diversity. It is up to the community to determine what this diversity looks like and what appropriate representation within the community looks like for both descriptive characterizations and decision-making bodies.

Free Prior Informed Educated Consent (FPIEC) Free prior informed educated consent must be established before any research is undertaken, at individual and collective levels, as determined by community governance structures. Prior informed consent is recognised as an ongoing process that is based on relationship; it should be reaffirmed and maintained throughout all phases of research. This principle recognises that prior informed consent requires an educative process that employs bilingual and intercultural education methods and tools, as appropriate, to ensure understanding by all parties involved. Establishing prior informed consent also presumes that all directly affected communities will be provided complete information in an understandable form regarding the purpose and nature of the proposed programme, project, study or activities, the probable results and implications, including all reasonably foreseeable benefits and risks of harm (be they tangible or intangible) to the affected communities. Indigenous peoples, traditional societies and local communities have the right to make decisions on any programme, project, study or activities that directly affect them. In cases where the intentions of proposed research or related activities are not consistent with the interests of these peoples, societies or communities, they have a right to say no.

Full Disclosure (FD) This principle recognises that Indigenous peoples, traditional societies and local communities are entitled to be fully informed about the nature, scope and ultimate purpose of the proposed research (including objective, methodology, data collection, and the dissemination and application of results). This information is to be given in forms that are understood and useful at a local level and in a manner that takes into consideration the body of knowledge, cultural preferences and modes of transmission of these peoples and communities.

Mindfulness (M) The concept of 'mindfulness' is an important value embedded in this Code, which invokes an obligation to be fully aware of one's knowing and unknowing, doing and undoing, action and inaction.

Ownership and Permission (O&P) This principle recognizes that communities are the rightful owners of their communal knowledge. Researchers will follow the processes set by communities to ask for permission to engage with the communities and their knowledge practices and for any reuse or redistribution of community knowledge. Communities may define their own permitted uses (such as for research, education, or personal uses only) or may define disallowed uses (such as commercial).

Precaution (P) This principle acknowledges the complexity of interactions, and thus the inherent uncertainty of effects due to research. The precautionary principle advocates taking proactive, anticipatory action to identify and to prevent harms resulting from research activities or outcomes. The prediction and assessment of such harms must include local criteria and indicators, thus must prioritize the voices of Indigenous peoples, traditional societies, and local communities. This also includes a responsibility to avoid the imposition of external or foreign conceptions and standards and to commit to the project for the duration of its investigation and reintegration of resulting benefits to the community.

Prior Rights and Responsibilities (PR&R) This principle recognises that Indigenous peoples, traditional societies, and local communities have prior, proprietary rights over, interests in and cultural responsibilities for all air, land, and waterways, and the natural resources within them that these peoples have traditionally inhabited or used, together with all intellectual property and traditional resource rights associated with such resources, knowledge, and their use.

Reciprocity, Mutual Benefit, and Equitable Sharing (RMBES) This principle recognises that Indigenous peoples, traditional societies, and local communities are entitled to share in and benefit from tangible and intangible processes, results and outcomes that accrue directly or indirectly and over the shorter and longer term for research and related activities that involve their knowledge and resources. Mutual benefit and equitable sharing will occur in ways that are culturally appropriate and consistent with the wishes of the community involved. Research partners are encouraged to consider the benefits they have received from community collaborators and the lessons that may be brought back to the academic community.

Remedial Action (RA) This principle recognises that every effort will be made to avoid any adverse consequences to Indigenous peoples, traditional societies, and local communities from research and related activities and outcomes. Not withstanding the application of standards set out by this Code of Ethics, should any such adverse consequence occur, discussion will be had with the local peoples or community concerned to decide on what remedial action may be necessary to redress or mitigate adverse consequences. Any such remedial action may include restitution, where appropriate and agreed.

Respectful Relationships (RR) This principle advocates for respectful, constructive relationships and acknowledges the roles, relationships and responsibilities each party has in the process of engagement. This includes the necessity for researchers to respect the integrity, morality and spirituality of the culture, traditions and relationships of Indigenous peoples, traditional societies, and local communities with their worlds. Research collaborations should also be aware of the power dynamics between outside researchers, community researchers, and community research participants and endeavor to approach heirarchical power structures in a culturally-appropriate manner.

Self-Determination (SD) This principle recognises that Indigenous peoples, traditional societies and local communities have a right to self-determination (or local determination for traditional and local communities) and that researchers and associated organisations will acknowledge and respect such rights in their dealings with local peoples and their communities. This right extends to decisions made about research topics and appropriate investigation methods when the research concerns the community and their knowledge.

Supporting Community Research (SCR) This principle recognizes and supports the efforts of Indigenous peoples, traditional societies, and local communities in undertaking their own research based on their own epistemologies and methodologies, in creating their own knowledge-sharing mechanisms, and in utilising their own collections and databases in accordance with their self-defined needs. Research with local communities should make efforts to center community perspectives and conduct such research in culturally relevant environments, rather than those that are convenient for outside researchers. Capacity-building, training exchanges and technology transfer for communities and local institutions to enable these activities should be included in research, development and co-management activities to the greatest extent possible.

The Dynamic Interactive Cycle (TDIC) This principle recognises that research and related activities should not be initiated unless there is reasonable assurance that all stages can be completed from (a) preparation and evaluation, to (b) full implementation, to (c) evaluation, dissemination and return of results to the communities in comprehensible and locally appropriate forms, to (d) training and education as an integral part of the project, including practical application of results. Thus, all projects must be seen as cycles of continuous and on-going communication and interaction.

Traditional Guardianship (TG) This principle recognises the holistic interconnectedness of humanity with the ecosystems of our Sacred Earth and the obligation and responsibility of Indigenous peoples, traditional societies and local communities to preserve and maintain their role as traditional guardians of these ecosystems through the maintenance of their cultures, identities, languages, mythologies, spiritual beliefs, internal diversity, and customary laws and practices, according to the right of self-determination. This principle also recognizes that communities are experiential and cultural experts of their traditional knowledge, and this knowledge is valid and legitimate without further mediation, translation, or interpretation.