Copyright & Fair Use

The purpose of the Copyright & Fair Use column is to keep readers informed on copyright as it affects the preservation and availability of historic recordings. We welcome your questions regarding copyright, and will endeavor to address them in these pages (we cannot, however, offer private legal advice). Comments and short articles describing your own experiences with copyright are also welcome. Please send submissions to Tim Brooks, Chair, ARSC Copyright & Fair Use Committee at tim@timbrooks.net. Opinions given here are those of the contributors. For general information visit the Committee's web page at www.arsc-audio.org and the site maintained by the Historical Recording Coalition for Access and Preservation, of which ARSC is a member (www.recordingcopyright.org).

On Rights Management in Anthropological and Linguistic Sound Collections

By Hugh J. Paterson III
Unaffiliated Collaborative Researcher

Archivists and curators benefit from well-documented rights declarations and agreements in the provenance of artifacts. Linguists, folklorists, anthropologists, and increasingly computer scientists, during the course of their work may create large and intricate collections of the soundscapes they encounter. Content may include personal narratives, word lists, audio books, talking dictionaries, and traditional music. This article looks at some of the legal issues encountered by academics as they produce audio and video recordings in minority language contexts.*

1 Introduction

This article seeks to expand the creator and curator's awareness of the scope and range of legal documentation needed for sound collections, so that archives can carry out their fiduciary responsibilities to preserve and grant access to the preserved content – especially within academic institutions. Given this context, I give preference to referencing academic publications rather than legal publications. However, this does not diminish the importance of legal publications such as laws, statutes, rulings, and opinions as primary sources.

Recent estimates indicate that there are about 7,000 spoken languages in the world without counting dialects. While the twenty most widely spoken languages all have over

80 million speakers, the average language community in Africa only has 30,000 speakers, and the average language community in the Americas has about 1,000 speakers. It is not uncommon for ethno-linguistic communities with small populations to view their heritage language and artifacts demonstrating the language in different ways than ethno-linguistic communities with large populations. In the contexts of academic collaborations to document small, often endangered languages, audio may be unpublished in the traditional sense of published audio such as music, radio productions, and more recently podcasts. However, this is not always the case as many formally published audio works have been released on the basis of anthropological field recordings (for example: Sapir 1965, Zemp 1971). Regardless of the formalities of the publication venue, the rules and laws of copyright equally apply. From a United States copyright law perspective (summarized in U.S. Copyright Office, Circular 1 2019), the application of copyright remains the same as in any other industry. Newman[1] (2007, 2011), writing from the perspective of U.S. law, lays out some considerations that linguists should take as they approach their work. Less well discussed are the kinds of resources archivists should request from linguists as they seek to document collections in archives or the kinds of documents which would be helpful to archivists if they were included in accessions in order to support the goals of Open Access collections. I wish to give an overview on the relevance of legal pluralism, or how working across jurisdictions can impact the perceived rights of possible claimants on the artifacts in a collection. Whereas publishing, in the formal sense, often carries with it an institutional process which addresses copyright, the more frequent case for the last decade has been to deposit collected materials in special "language archives"[2]. Language archives in many cases are smaller organizational units within (academic) institutional libraries or public institutions (museums). However, in some cases they may be private collections, part of an NGO or other corporation, or part of a community heritage center. Many language archives are aware of various social sensitivities related to specific collections, due to academic discipline norms of discussing use-cases where the socio-cultural sensitivities of the persons recorded are explained and respected. However, the issue of copyright is one of ownership – specifically of economic rights – not one of respecting cultural norms[3]. Copyright is therefore often an inadequately documented issue, even at language archives. In addition to language archives, it not uncommon for special collections of university libraries to also contain sound recordings as supplementary materials to associated dissertations and theses. In these cases, materials are stored and made accessible, sometimes without any guidance on possible rights considerations. With the abundance of speech collections in different languages and the large diversity of institutions supporting the access and preservation of these materials, broader communication about the kinds of curation these collection require is needed. Within the context of this broader communication, it is important to emphasize that not all content in under-documented languages has the same sensitivities as is frequently discussed in the language documentation literature. For example, some communities have taboos about hearing the voices of deceased persons, while others have taboos about non-community people hearing or experiencing rites and ceremonies.

Good rights management will document: Cases of legal plurality, cases of conflicting approaches to concepts of ownership, the legal framework under which rights are asserted, contracts engaged in when the recording was conducted including any work-for-hire arrangement for the recorded artist and the academic (producer/sound engineer), informed consent documentation, institutional review board agreements, any national research visa

permissions, consent and wavers related to the European Union's *General Data Protection Regulation* (GDPR)[4], as well as any permissions or restrictions for creating derivative works including the use of artificial intelligence and machine translation to create language models or transcriptions. In the rest of this article I focus on the issues of legal plurality and conflicting concepts of ownership, but first some background information.

2 Preliminaries

Artifacts and Open Access in the sciences

Internationally, agencies funding language documentation and linguistic work require grant applicants to provide "data plans" including where they are going to archive their funded creations. It is also often a stipulation of public funded grants that data generated (created artifacts) be publicly accessible[5]. Across the academy, data is coming under scrutiny with regard to its status as being *scientific*. That is, the results of an academic endeavour need to be examinable and repeatable – regardless of the success of the scientific experiment. Broadly across the sciences and academic data management (including archival activities), this is discussed under the label *FAIR Data Principles*. FAIR stands for Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). FAIR builds upon the ideals articulated in 2010 under the term Open Access (OA) in the Budapest Open Access Initiative[6,7]. To the best of my knowledge, the exact nature of how OA plays out with copyright is not well-litigated or exposited in the academic legal journals[8]. As a general principle of the OA approach to science and the ecology of academic works, copyright is retained by an author and a work is licensed using a *Creative Commons with Attribution 4.0* license, or if the work is software, a *GNU* or *MIT* style license. Data is licensed under a *Creative Commons CC0* style license also known as a public domain dedication[9]. Within the sciences, it is not uncommon to call the product of research "data". This terminology has sometimes been used in the linguistic literature[10]. However, as an archivist, I prefer to call recordings: "artifacts" or "creative works". Often linguists will reference a variety of artifacts and use them as evidence in their academic publications, and hence discuss them in the same formulaic rhetoric that other scientists will use to discuss other kinds of observations, e.g., chemical interactions or temperatures.

Copyright and freedoms of use

Within the commercial and for-profit recording industry, copyright is only one of several legal issues which recording artists should address, e.g., is the artist under exclusive contract with another studio? The academic recording context can be equally complex and must be duly considered. The goal of recordings created under the umbrella of academic activity generally has a distribution model and an intended audience distinct from major publishers of other kinds of audio artifacts. Broadly there are two categories of recordings in the domain of linguistics. First, there are those of an experimental nature which have some sort of investigative purpose or structure which binds them together for contrast and comparison. Second are those whose content is more akin to folklore and oral history recordings. They are demonstrative of "speech styles" and "subject matter" discussed in the language community[11]. Academics face pressures from funders to release both types of recordings within OA frameworks. An additional pressure facing academics is related to their career standing and career prospects. That is, academics are concerned with how

many times their work is referenced by other scholars. Open Access materials are argued to be more available and therefore more likely to be reused. One of the requirements of OA materials is that licenses must have overt clauses enabling reuse. This precludes "implied use" licenses. Some organizations do not apply any license to copyrighted content but still make it available online. In the U.S. this is known as an "implied use" license[12]. In contrast to "implied use" licenses, Creative Commons licenses directly address the reuse and derivative works issues. Creative Commons licenses have their merit on the basis of copyright law (Lessig 2004). They actually use copyright law to give freedoms of use – they don't absolve or do away with copyright. This seems to be a common mis-understanding about the use of Creative Commons licenses among academic linguists (though there seem to be other common misconceptions such as non-profit organizations should use the Creative Commons license with the Non-Commercial clause). With OA it is still important to answer the question: is copyright the only legal limiter in the use of an artifact? This question is part of the framing of the first two points in the *Legal Interoperability of Research Data: Principles and Implementation Guidelines* (Agosti et al. 2016) where the discussion concerns all rights not just copyright[13]. So, archivists and creators both need to ask questions like: what sorts of documentation support and defend the Open Access principles and language focused artifacts? And what sort of documentation challenges need to be overcome to make linguistic data as reusable as possible, according to the intent of the collection?

Copyright and legal strategy

When discussing copyright and other legal issues, four things are helpful to keep in mind. First, what the law *says* and *when or how it is applied* are two separate things. Second, nobody knows what the law *means* (regardless of what it *says*) until after a judge applies the law to a case. Third, it takes a lot of money to bring a case before a judge to solve issues of rights ownership. Fourth, the dispute is not over the ownership of the artifact, but rather the rights related to the artifact. Unfortunately, trust in many interpersonal and inter-organizational relationships is often broken long before the time that a judge makes the law clear in a particular case. This broken trust may divide communities for generations. Within the context of discovering what the law says and what it means and when and where it is to be applied, individuals and organizations craft strategies for engaging with copyright law.

Copyright litigation related to audio as artifacts of research is infrequent. Academic disagreements rarely make it to court. Continuous legal conflict for academic institutions has a tendency to lower the value of their brand, a core income driver. Of those cases which I have found that have made it to court on issues discussed herein, none are specifically about audio artifacts. As such the relevance of these cases is only applicable to audio artifacts by way of analogy, a common tool in law when no clear precedent is found.

People who are concerned with copyright are often looking to shield themselves or their organizations from liability, and so understanding the law is perceived as important for charting a path which avoids liability. Of course this includes the costs of litigation, damages, and the public shame which comes with controversy. From a business strategy perspective, this view esteems intellectual property, including copyrights, as a sparse resource and connects economic models to the delivery of this sparse resource to interested

parties. In contrast to the well charted path seeking to avoid liability, knowing the law can be perceived to be an asset by a party who desires to assert sovereignty and also control the use of artifacts. Artifacts are powerful tools in crafting socio-political narratives and creating economic opportunities. We might be tempted to conceive of economic opportunities as the ability to market Nashville's music[14] via Spotify, Grooveshark, Last. fm, SoundCloud, or a host of other audio distributors and record labels. However, I would like to draw attention to other economic opportunities which should be considered within the context of copyright discussions. For example copyright protects the economic interests of John Wayne films, *Davey Crockett: King of the Wild Frontier* (Disney), episodes of *The Lone Ranger* TV series, and Lucky Luke comics. Each of these sets of media, but not just these media (Baigell 1990), portray indigenous Americans within a narrative or context which justifies their subservient position to people of European decent. The media is in many respects an articulation, a side narrative to, or an evolution of the "manifest destiny" ideology which favors economic opportunities for capitalistic Europeans at the expense of the ethnics whose documented existence on the land predates the western expansion of the United States by hundreds of years – even the establishment of the U.S. constitution. It is hard to ignore the capitalistic heritage of the U.S. as the colonies were themselves constructs of capitalistic ventures. Copyright law within the U.S. has its heritage in this capitalistic perspective – framed within an adversarial legal system. So, copyright – the ability to exploit a creative work, to choose where or how it is contextualized – becomes an important issue for many indigenous communities. When it comes to language related artifacts, and especially audio artifacts actual case law in the U.S. is sparse[15]. However, the academic literature does contain several discussions related to language materials.

Hinton and Weigel (2002:168–170) discuss a case with an un-named U.S. tribe where a linguist was working with two tribal members on a dictionary. Prior to publication a review committee was formed with other tribal members who reviewed the work. The resulting review process involved the contribution of many new additions and modifications, for which the members of the review committee felt they were not adequately acknowledged. The discussion ended up centering around copyright which was set to be in the linguist's name. The dictionary was published but at the cost of the relationship between the linguist and the language community. A second example in the same volume (Hill 2002) discusses how the University of Arizona Press negotiated copyright and distribution of the Hopi dictionary with the Hopi tribe. The preserved relationship eventually led to a second printing of the dictionary.

Other academic literature does approach the issue of copyright from a broader perspective and includes discussion of visual media. Brown (1998) frames his critical question *can culture be copyrighted?* as a response to actions taken by Harvard's Peabody Museum regarding visual media in their holdings[16]. The museum's position and actions as articulated in Sandager (1994) were to seek the advice of Navajo consultants concerning the restoration of images within the collection of materials provided by A. M. Tozzer. The works in question had previously been published in a volume, but the hand sketched images were of "earth images", hand drawings traditionally destroyed at the completion of Navajo healing rituals. Brown situates his discussion in the context of a preservation organization's fiduciary responsibility to those who bequeath collections to the organization and relationships with the descendants of those who might have shared their culture with collection creators, e.g., Tozzer. Meta-context of the cultural documentation activity

is vitally important for this singular reason: When both the parties of a documentary endeavour (the documenter and the documented) are deceased, how shall an archive respond to various kinds of claims about the artifacts?

Notice that in each of these cases within the North American context, the critical issue as framed in publications is not the copyright of the language or the language materials per se, but rather the power to craft the narrative which presents the materials and the people they represent.

The desire to assert sovereignty over artifacts can be found among some ethno-linguistic communities, primarily those who have a strong material culture regarding the objects which represent their cultural heritage. This position is justified given the socio-historical context. However, it stands in contrast to those who form strong material bonds with artifacts for the purpose of financial exploitation of those artifacts, e.g., recording artists or preservation organizations whose funding streams are dependant on the artifacts they hold. From a business strategy perspective, Open Access is equally about the ability to create economic opportunities. However, proponents of Open Access have generally adopted a business model where the artifact no-longer represents the strategic economic opportunity. Rather the opportunity, and the related business advantage, is secured by maintaining a team which can innovate in creative ways to adapt artifacts and to create more artifacts. Even within the academy, the goal in progressive universities is not to hold knowledge, but to facilitate the discovery of new knowledge through strategic industry collaborations (see discussion in Kelli et al. 2013). In this way, sharing knowledge under Open Access creates competition between teams and encourages the discovery of new knowledge via innovation – it positions the academic institution as a knowledge discovery service and an innovation service within a larger social enterprise. Open Access takes the focus off of the artifact and places it on the process to create the artifact and what can be created using the artifact. Language archives can ask the following critical question: Are we positioned to guard artifacts or are we engaged in facilitating social innovation through the exposure of the knowledge we preserve?

The pursuit of economic opportunities from engaging with language communities is not lost on missionaries. For example, Wells (1977) discusses the history of the prominence of African masks in art collections, attributing missionary Dr. George Harley with creating the African mask art market by selling masks to collectors and museums such as the Harvard's Peabody Museum. Harley's practice opened him up to questions of exploitation for personal gain, at the expense of the communities he was purporting to serve. A similar question arises with the re-purposing of the Bible (both in audio formats and textual formats) for machine translation and speech-to-text resources. The use of audio versions of the Bible (because the Bible has been translated into hundreds of languages) is increasing in popularity within academic research (Gauthier et al. 2016, Black 2019, Zanon Boito et al. 2020). It remains to be seen if Bible translation organizations will capitalize on their claimed copyrights on Bible translations – even if they attempt to implement "AI for Good". Both machine translation and speech-to-text (also know as *automatic speech recognition* or ASR) are examples of *arti icial intelligence* (AI), a broad cover term for a variety of technical processes related to pattern matching. The basic process involves using a set of audio or text language-resources and then passing them through an AI tool to create a language based "model" and then applying the model within a software process to generate output (transcription or translation) from data the model has never previously encountered.

The prevailing thought has been that the original data used to create the language model, the language model, and the outputs are all separate works, governed and subject independently to rights frameworks such as copyright, neighboring rights, privacy rights, moral rights, etc. (Kelli et al. 2020a, 2020b, and also discussion in Klavan et al. 2018, Kelli et al. 2019 and Kelli et al. 2019). However at the moment, I remain unconvinced that language models are independent works from their training data, rather I see them as derivative works. For example, copyright covers sculptures and their forms. If an artist takes someone else's sculpture and creates a mold for it and then casts new material in the mold, that would be copyright infringement because the shape of the original mold was used. Essentially language models do the same thing, they take a mold of the language data by abstracting the shape of the data. The critical difference is claimed to be that the training data is not recoverable from the "mold". However, this doesn't negate the fact the the training data was used to create the model and that without the training data the model would not be what it is. That is, the essence of the AI language model is directly related to the content and "data shape" of the training data. The extent that a language model is to be considered a new expression of art, rather than a derivative work remains to be clarified until a judge decides in case law. Many in the business world praise the advances of AI technology. More simply put it is the commercialization of pattern matching. So when companies exploit the use of pattern recognition software, it becomes an ethical issue on which patterns they are looking for and what sort of biases are contained in the training data. Rudin (2019), along with others, has raised ethical questions on the use of AI tools and argues for transparent training data along with transparency in the ability to determine how data is processed throughout the pattern matching process. If language models are derivative works, this could be an infringement of moral rights or other inalienable rights because many Bible recordings do not indicate the rights of the speakers/voice actors.

3 Plurality and con licting approaches to ownership

Within language research contexts, the creation of audio artifacts often means crossing national borders, meaning that legal systems from one or more countries will apply. This is plurality – the impact of two or more legal systems on the creation process. It also means working across cultures and with people (who from an recording industry perspective fill the roles of *author* and/or *recording artist*) who might not be accustomed to thinking about the created artifact in the context of a legal framework. This might be the legal framework of their own country or another country. For example, in the U.S. legal framework several different parties may have copyright claims on a work for various types of contributions. For instance, the lyrics of a song, the music of a song, the composition of a song, and the performance of a song may all be copyrighted by separate entities. Even if a song's lyrics and composition are legally public domain, the manifestation of that work as performed by a particular artist is covered by copyright and usually assigned to that artist. As we look at the kind of works recorded in linguistic research and language documentation activities, a folktale which is widely know, having been in circulation for generations, is likely not copyrightable at the content level; however, a performance of that folktale is copyrightable. The various natures of copyright claims made on a work are more familiar to those who work in the commercial recording industry, e.g., *recording engineers* or *producers*. However,

in linguistic research, people primarily identify with roles like *graduate student*, *primary investigator*, or *professor*. This often leaves the various relationships between rights holders and content unexplored. It doesn't mean that a legal framework doesn't apply to the artifact, only that it is not the primary framework under which trust is built between the parties creating the artifact[17].

In certain cultures once there is trust to co-create something (including audio artifacts), there is an implicit understanding about how the artifacts will be used. In contrast to these unspoken and assumed norms, legal frameworks like U.S. law prescribe a best practice of making explicit agreements in writing before recording begins (for discussion see van Driem 2016). Even within U.S. academic circles, primary consideration is not legal but ethical. In my experience, most arguments for policy change at institutional levels stem from ethical arguments rather than legal arguments. In a similar vein, legal changes at the national level (in the U.S.) usually stem from economic arguments, rather than clarifying language on the basis of judicial rulings. A further consideration in this regard is that I have yet to see an institutional review board approved "informed consent" statement in linguistic research address the issue of copyright. Informed consent is generally treated like a license with specific use outlined, totally side-steps the issue of ownership, and situates the artifact in a framework of possesorship. The framework of possesorship regardless of the validity or application of any copyright claims remains a necessary framework for archives. It gains additional prominence when considering content which has been placed in the public domain or whose copyright term has expired. Possesorship of digital artifacts, which might have several copies, is unlike possesorship of physical artifacts which usually only have a single copy[18].

Geographic complexity

Geographic complexity is an issue to consider because often times the location and country where an artifact is created is not the same country as where the artifact preservation activities occur. Cultural issues aside, my experience has been that recording engineers (in a loose sense of *engineer*) have mostly approached the activity as a personal matter between two people and ignored national laws in the context where the research is conducted. When copyright is assumed to protect the recordings, it is assumed to be the copyright framework of the country doing the preserving activity, often the same country the recording engineer/researcher/linguist/professor is based in, rather than the country in which the recording activity was conducted. This can make a difference as copyright durations vary from legal framework to legal framework, e.g., some countries may lack a public domain for cultural heritage materials. National legal frameworks may have far reaching implications in the non-copyright rights associated with research artifacts. For instance, civil law frameworks (France) may prescribe different rights to creative works (artifacts) than common law frameworks (USA). This could impact an American linguist making language or music recordings in Francophone Africa (which generally follows the French legal framework), whereas the U.S. legal system does not acknowledge moral rights[19]. Another example might be the privacy rights which are prescribed by the GDPR and apply to artifacts when Europeans create these artifacts even as they conduct research outside of the European Union. Privacy rights are a new area of legal practice with great financial promise for legal experts considering the existence of the GDPR, Brazil's privacy framework *Lei Geral de Proteção de Dados*, and California's privacy framework.

Legal pluralism

Legal pluralism is the concept that two legal systems both have relevance in a social setting. There are a variety of contexts where we might encounter legal pluralism. The first might be a conflict between a state (sub-national) and a federal (national) statute or national law and Sharia law as practiced in Nigeria or Malaysia. There are generally procedures for resolving the sorts of conflicts where an acknowledged hierarchy exists between the legal entities. However, the situation is more complicated in contexts where there there is no legal agreement between the entities, such as in cases of national law and Sharia law as practiced in London, UK. One way to look at these sorts of conflicts is that they represent two independent social contracts within the same geographic space. It is not too far a jump to call social contracts "culture". Cultural norms change, but documenting the agreements under which recordings were conducted can help to establish the legitimacy of artifacts and the legitimacy of their continued use by a variety of stakeholders, even when cultural norms change and the artifacts do not[20]. An example of this might be in cultures where women have different socio-cultural liberties than men do. For instance, a married woman may not be free to offer a recording of a song sung only by women with out the permission of her husband. In this way there is not an express legal pluralism but rather a cultural perception that the "rights holder" is not the person actively involved in the recording. This is not necessarily unlike a recording artist having an exclusive contract with particular label. The label's representative must give consent for the artist to conduct a recording outside of their defined relationship. With urbanization impacting the cultures of many formerly rural ethnic communities, many of these cultural norms are evolving. Today's norm is tomorrow's "historical custom". While national law might expressly indicate who is the copyright holder, it is another, but relevant matter as to who is granting the use of the artifact or the permission for a person to be involved in the creation of the artifact.

Another issue in legal pluralism is sometimes the notion that an indigenous tribe or community has a different set of customs than what is prescribed by national laws. Some legal theorists seek to codify these social contracts within nationally recognized legal frameworks, while others argue for a *sui generis* approach[21]. As is mentioned by Brown (1998) and Tatsch (2004) sometimes the idea is put forward that a language is owned by "the community"; and therefore, "the community" in some way has rights to the artifacts which contain the language. Janke (2016) points out these types of social-ownership arguments have failed in Australia. However, when a legal strategy was chosen where it was argued that a particular manifestation was copyrighted by the individual, that upheld in Australian court. The arguments for "community ownership" see challenges because copyright assignment requires either a person or a registered entity. A "language community" is neither of these[22]. As Widlok (2013) points out, defining "the community" is hard at best and functionally impossible for archives to implement. However, countering these arguments from a legal point of view is challenging and can be relationally damaging – a socially sensitive issue. Arguments for the amorphous "language community" are most often seen in Australia, Canada, and the United States: all locations where laws were historically used to limit opportunities for indigenous people and locations where language is a highly regarded component of community identity and cultural heritage.

The issue of legal ownership rights has been muddied by terminology choices at archiving institutions by the use of the term "access rights". Access methods, permissions,

and privileges are best kept separate from the issue of rights related to ownership and the exercise of rights granted by copyright. In contrast to access methods, permissions, and privileges granted by contracts or institutions, *rights* are granted by a government via statutory law[23]. It is conceivable that a court could limit a copyright holder's ability to restrict access to their materials, which is a limit on the full exercise of copyright. Such a limit would impact access, but framing this situation as a "right to access" is different than framing it as a "limit on copyright", although the functional impact might be the same.

4 Conclusion

The purpose of the current article is to point out some extra-copyright issues and some ways that sound artifact creators and archivists can support each other in the use and understanding of collected artifacts. Janke (2016) points towards protocols used in Australia which functionally force researchers to engage in rights management issues. This is done by vesting copyright in the individual being observed, rather than the individual doing the observing. The Australian example works in Australia where there is a clear social understand of how the protocols are helpful and usable by both the observed and the observer. My experience working with minority communities in Nigeria and Mexico are not nearly as complicated as the Australian situation. However, in any research context, researchers can always pursue a position where they license the artifact from the observed rather than owning it outright. By using standardized licenses like Creative Commons licenses, the research artifact can have a broad range of usage. However in the European Union and other jurisdictions with privacy protection schemes, it remains to be seen if Creative Commons licenses are sufficient to indicate that data also contains waivers of any privacy based rights, not granted on the basis of copyright protections.

References

Agosti, Donat, Enrique Alonso Garcia, Baden Appleyard, Christoph Bruch, Robert Chen, Gail Clement, Willi Egloff, Herbert Gruttemeier, Simon Hodson, Maria Llosent, J. Bernard Minster, and Paul F. Uhlir. 2016. *Legal Interoperability of Research Data: Principles and Implementation Guidelines.* Paul F. Uhlir, Enrique Alonso Garcia, and Robert Chen, eds. (RDA-CODATA Legal Interoperability Interest Group.) Belgium: Research Data Alliance. doi:10.5281/zenodo.162241

Anderson, Jane. 2005. Indigenous Knowledge, Intellectual Property, Libraries and Archives: Crises of Access, Control and Future Utility. *Australian Academic & Research Libraries* 36(2). 83–94. doi:10.1080/00048623.2005.10721250

Ashmore, Louise. 2008. The role of digital video in language documentation. *Language Documentation and Description* 5. 77–102. http://www.elpublishing.org/PID/064

Atkinson, Karen J. and Kathleen M. Nilles. 2008. *Tribal Business Structure Handbook.* Washington, D.C: The Office of the Assistant Secretary – Indian Affairs U.S. Department of Interior. https://www.irs.gov/pub/irs-tege/tribal_business_structure_handbook.pdf

Baigell, Matthew. 1990. Territory, Race, Religion: Images of Manifest Destiny. *Smithsonian Studies in American Art* 4(3/4). 3–21. https://www.jstor.org/stable/3109013

Black, Alan W. 2019. CMU Wilderness Multilingual Speech Dataset. Paper presented at:

ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom. doi:10.1109/ICASSP.2019.8683536

Brown, Michael F. 1998. Can Culture Be Copyrighted? *Current Anthropology* 39(2). 193–222. doi:10.1086/204721

Fitzgerald, Anne, Neale Hooper and Brian Fitzgerald. 2010. Enabling Open Access to Public Sector Information with Creative Commons Licences – the Australian Experience. In Brian Fitzgerald, *Access to Public Sector Information: Law, Technology and Policy,* 71–138. Sydney, Australia: Sydney University Press. https://eprints.qut.edu.au/29773/

Gauthier, Elodie, Laurent Besacier & Sylvie Voisin. 2016. Automatic Speech Recognition for African Languages with Vowel Length Contrast. Paper presented at: 5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU). Yogyakarta, Indonesia. doi:10.1016/j.procs.2016.04.041

Hill, Kenneth C. 2002. On Publishing the *Hopi* Dictionary. In William Frawley, Kenneth C. Hill and Pamela Munro, *Making dictionaries: preserving indigenous languages of the Americas,* 195–218. Berkeley, California: University of California Press.

Hinton, Leanne and William F Weigel. 2002. A Dictionary for Whom? Tensions between academic and Nonacademic Functions of Bilingual Dictionaries. In William Frawley, Kenneth C. Hill and Pamela Munro, *Making dictionaries: preserving indigenous languages of the Americas,* 155–70. Berkeley, California: University of California Press.

Janke, Terri. 2016. Ensuring Ethical Collaborations in Indigenous Arts and Records Management. *Indigenous Law Bulletin* 8(27). 17–21.

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Ramūnas Birštonas, Penny Labropoulou, Kadri Vider, Irene Kull, Gaabriel Tavits, Age Värv and Vadims Mantrovs. 2020a. Impact of Legal Status of Data on Development of Data-Intensive Products: Example of Language Technologies. In Kiril Simov and Maria Eskevich (eds), *Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019.* 69–74. CLARIN. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-100814

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Ramūnas Birštonas, Penny Labropoulou, Kadri Vider, Irene Kull, Gaabriel Tavits, Age Värv and Vadims Mantrovs. 2020b. Impact of Legal Status of Data on Development of Data-Intensive Products: Example of Language Technologies. In A Damberga, *Legal Science: Functions, Significance and Future in Legal Systems II – Collection of Research Papers in Conjunction with the 7th International Scientific Conference of the Faculty of Law of the University of Latvia (16–18 October 2019, Riga),* 383–400. Riga, Latvia: University of Latvia press. doi:10.22364/iscflul.7.2.31

Kelli, Aleksei, Krister Lindén, Arvi Tavast, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits and Age Värv. 2019. The Extent of Legal Control over Language Data: the Case of Language Technologies. *Proceedings of CLARIN Annual Conference 2019.* 69–74. Utrecht, Netherlands: CLARIN ERIC. https://researchportal.helsinki.fi/en/publications/the-extent-of-legal-control-over-language-data-the-case-of-langua

Kelli, Aleksei, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavrilidou and Pavel Straňák. 2019. Processing personal data without the consent of the data subject for the development and use

of language resources. In Inguna Skadin and Maria Eskevich (eds), *Selected papers from the CLARIN annual conference 2018, Pisa, 8–10 October 2018.* Linköping Electronic Conference Proceedings №159:72–82. Linköping, Sweden: Linköping University Electronic Press, Linkpings universitet. https://epublications.vu.lt/object/elaba:40541287

Kelli, Aleksei, Tõnis Mets, Lars Jonsson, Heiki Pisuke and Reet Adamsoo. 2013. The Changing Approach in Academia-Industry Collaboration: From Profit Orientation to Innovation Support. *Trames. Journal of the Humanities and Social Sciences* 17(3). 215–41. doi:10.3176/tr.2013.3.02

Klavan, Jane, Arvi Tavast and Aleksei Kelli. 2018. The Legal Aspects of Using Data from Linguistic Experiments for Creating Language Resources. In K Muischnek and K Müürisep, eds. *Human Language Technologies – The Baltic Perspective,* 71–78. (Frontiers in Artificial Intelligence and Applications), 307. Amsterdam, Netherlands: IOS Press. http://ebooks.iospress.nl/volumearticle/50306

Laitin, David D. 2000. What Is a Language Community? *American Journal of Political Science* 44(1). 142–155. doi:10.2307/2669300

Lessig, Lawrence. 2004. The Creative Commons*. *Montana Law Review* 65(1). 1–14. https://scholarship.law.umt.edu/mlr/vol65/iss1/1

Margretts, Anna and Andrew Margretts. 2012. Audio and Video Recording Techniques for Linguistic Research. In Nicholas Thieberger, *The Oxford Handbook of Linguistic Fieldwork,* 13–53. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199571888.013.0002

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

Newman, Paul. 2011. Copyright and Other Legal Concerns. In Nicholas Thieberger, *The Oxford Handbook of Linguistic Fieldwork,* 430–456. Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/9780199571888.013.0020

Patrick, Peter L. 1999. The Speech Community: Some Definitions. Paper presented at: NWAVE-28, 17 October. Toronto, Canada. https://web.archive.org/web/20130927110820/http://orb.essex.ac.uk/lg/lg232/SpeechComDefs.html

Patrick, Peter L. 2008. The Speech Community. In J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes, *The Handbook of Language Variation and Change,* 573–97. Malden, MA: Blackwell Publishing Ltd.

Rudin, Cynthia. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5). 206–215. doi:10.1038/s42256-019-0048-x

Sandager, Elizabeth. 1994. Ethical implications of the documentary record. *New England Archivists Newsletter* 21(2). 4–6.

Sapir, J. David. 1965. *Diola-Fogny, Recorded By J. David Sapir – The Music Of The Diola-Fogny Of The Casamance, Senegal [Vinyl, LP, Album].* New York, NY: Folkways Records – FE 4323.

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi

Ngwabe Neba, Sebastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort and Tony Woodbury. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation* 13. 545–563. http://hdl.handle.net/10125/24901

Suber, Peter. 2008. Gratis and libre open access. *SPARC Open Access Newsletter* 124. https://dash.harvard.edu/handle/1/4322580

Suber, Peter. 2012. The rise of libre open access. *SPARC Open Access Newsletter* 164. https://dash.harvard.edu/handle/1/32989158

Tatsch, Sheri. 2004. Language Revitalization in Native North America – Issues of Intellectual Property Rights and Intellectual Sovereignty. *Collegium antropologicum* 28(1). 257–262. https://hrcak.srce.hr/27951

U.S. Copyright Office. 2019. *Copyright Basics.* (Circular №1.) Washington, D.C: U.S. Copyright Office. https://www.copyright.gov/circs/circ01.pdf

van Driem, George. 2016. Endangered Language Research and the Moral Depravity of Ethics Protocols. *Language Documentation & Conservation* 10. 243–252. http://hdl.handle.net/10125/24693

Weitzmann, John Hendrik and Philipp Otto. 2011. *Term Extension for Related Rights in Sound Recordings.* Berlin: iRights. https://irights.info/wp-content/uploads/userfiles/Schutzfrist_A5_engl_final.pdf

Wells, Louis T. 1977. The Harley masks of northeast Liberia. *African arts* 10(2). 22–27, 91–92. doi:10.2307/3335179

Widlok, Thomas. 2013. The Archive Strikes Back: Effects of Online Digital Language Archiving on Research Relations and Property Rights. In Mark Turin, Claire Wheeler and David Nathan, eds. *Oral Literature in the Digital Age: Archiving Orality and Connecting with Communities.* (World Oral Literature Series №2, Open Book Publishers. doi:10.11647/OBP.0032.03

Wilkinson, Mark D., 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. 't 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). 160018. doi:10.1038/sdata.2016.18

World Intellectual Property Organization (WIPO), International Labour Organization (ILO) and the United Nations Educational, Scientific and Cultural Organization (UNESCO). 1961. *Rome Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations.* Geneva, Switzerland: World Intellectual Property Organization (WIPO). https://www.wipo.int/treaties/en/ip/rome

World Intellectual Property Organization (WIPO), International Labour Organization (ILO) and the United Nations Educational, Scientific and Cultural Organization (UNESCO). 1981. *Guide to the Rome Convention and to the Phonograms Convention.* (WIPO Publication №617(E).) Geneva, Switzerland: World Intellectual Property Organization (WIPO). https://www.wipo.int/edocs/pubdocs/en/copyright/617/wipo_pub_617.pdf

Endnotes

1. Paul Newman holds both a Ph.D. in Linguistics and a Juris Doctorate.

2. For a quick list of language archives consider the archives which are part of the Open Languages Archive Community: http://www.language-archives.org/archives.

3. From page two of the U.S. Copyright Circular №1, copyright provides the owner of copyright with the exclusive right to: Reproduce the work in copies or phonorecords. Prepare derivative works based upon the work. Distribute copies or phonorecords of the work to the public by sale or other transfer of ownership or by rental, lease, or lending. Perform the work publicly if it is a literary, musical, dramatic, or choreographic work; a pantomime; or a motion picture or other audiovisual work. Display the work publicly if it is a literary, musical, dramatic, or choreographic work; a pantomime; or a pictorial, graphic, or sculptural work. This right also applies to the individual images of a motion picture or other audiovisual work. Perform the work publicly by means of a digital audio transmission if the work is a sound recording. Copyright also provides the owner of copyright the right to authorize others to exercise these exclusive rights, subject to certain statutory limitations.

4. http://data.europa.eu/eli/reg/2016/679/2016-05-04

5. The exact nature of ownership for publicly funded projects which create copyrightable works varies. For instance, in the United States, it is often the case that creations by the federal government are in the public domain. However state and local governments often retain copyright, or charge a service fee for access to government resources. Additionally, resources created by contractors to the government or grantees of government funds may or may not be placed in the public domain depending on the circumstances. An American's perspective that government resources are in the public domain does not translate to Australian, New Zealand, or UK jurisdictions which by default retain copyright over their creations. However, there is a social movement to use Creative Commons licenses with government created works to enhance access and reusability of these works (Fitzgerald et al. 2010).

6. https://www.budapestopenaccessinitiative.org

7. For a fuller discussion of OA in Documentary Linguistics see Seyfeddinipur et al. (2019).

8. Open Access's interplay is with copyright is no doubt related to its reliance on Creative Commons licenses and their ability to withstand litigation.

9 Artifacts which are eligible for copyright may be released under any license at any time. Because Creative Commons licenses are non-revocable their duration is until the end of the copyright period. For example, a sound collection recorded in 1981 might have been recorded on magnetic tape. A subsequent transfer to a digital format might provide opportunity for new copyright claims of over the digital manifestation, but the original copyright claims still apply to content and composition. Provided all rights holders are in

agreement, the digital manifestation could be licensed using Creative Commons at any time in the future and last the duration of the copyright period. Open Access then comes into play when an organization makes a commitment to provide digital access to that artifact without any encumbrance.

10. While most linguists and archivists consider audio artifacts to meet thresholds required for copyright, some linguists make the claim that audio artifacts are by nature "data". This might open up the possibility of interpreting them as factual rather than artistic works. As far as I know this line of reasoning has not been explored legally. If linguistic recordings are "data", this may impact their use and dissemination under fair use. See argumentation in *The Swatch Group Management Services Ltd., vs. Bloomberg L.P.*, No. 12-2412–cv, 12-2645–cv (United States Court of Appeals for the Second Circuit January 27, 2014). In that case, Bloomberg argued that dissemination of a copyrighted file for the public news was fair use. (They also argued that the file was data.) It might be a far fetch to consider language archives as analogous to publicly traded companies and their content as newsworthy for the public good. However, there are likely dozens of linguists who are eager to consume the latest "news" or linguistic "data" on any given language.

11. Documenting the performing arts in endangered language communities is also important and actively being done. Documentation is encouraged in video formats, as the documentary record is richer and contains body position, gesture, some of the situational context, as well as the auditory communication component (Ashmore 2008, Margretts and Margretts 2012). Ensemble and solo performances are also within the purview of language documentation and culture documentation activities. While the current article primarily discusses audio artifacts, rights considerations should be evaluated for performances when it is clear that a performance was recorded. The threshold for performance may be different in different jurisdictions, further complicating the matter of facilitating broad discussion among legal professionals, archivists, and language documentation practitioners. In contexts where the Rome Convention (1961 and as amended) applies, there are "neighboring rights", also known as "related rights", to be considered. These are the rights of the performers in the recorded performance. The U.S. is not signatory to the Rome Convention and so does not acknowledge these rights, nor do other countries acknowledge these rights for U.S. performers. How and when "neighboring rights" apply in language documentation contexts, as far as I know, has yet to be explored. For further discussion of "neighboring rights" see the *Guide to the Rome Convention and to the Phonograms Convention* (1981) and Weitzmann and Otto (2011).

    A second point under the Rome Convention relates to archives and the Open Access content they distribute. The question is: *are archives broadcasters?* As far as I know, no legal theory has been put foreword suggesting digital data repositories are better considered as broadcasters than archives. So, if this affords archives any more rights or protections than they already have remains to be seen.

12. In the past there have been some variations on the implementation of OA. Some terms used to describe those include *Libre OA* and *Gratis OA*. The differences primarily relate to what one may do with the accessed materials — one is limited to access while the other has overt reuse permissions (Suber 2008, 2012).

13. There are two things worth pointing out. First, Agosti et al. (2016) was published prior to the enactment of the EU's GDPR. Second, the recommendation is that there should be a preference for licensing data with a public

domain dedication such as CC0. Within linguistics, I have rarely seen the use of CC0 within academic data, though I suspect that this is not for legal reasons but rather for the following two reasons. First, moral rights can not be asserted with CC0 and in many cases the association of a name with the data is not just important to minority language speakers but also to academics. Second, Creative Commons with Attribution 4.0, encapsulates a requirement for attribution. Attribution via referencing is important to the popularity contest which drives academic job security. Neither of these issues is really a copyright issue per se, rather one is a moral rights issue and the other a matter of academic professionalism. It is possible that biases in academic circles lead to the unnecessary use of copyright as a legal instrument.

14. Truly music comes from all corners of the globe, but Nashville is an acknowledged center of music production.

15. If a reader knows of specific legal cases, please send me an e-mail.

16. Though this example is not specifically about language materials, language materials are included in the kinds of materials that Vernon Masayesva, chairman and CEO of the Hopi tribe, requested from museums in 1994—though the main thrust of the letter seems to be over concerns about burial practices.

17. Traditionally, it has been understood that a researcher might be from a university and would not be a member of an ethno-linguistic minority population. However, these assumptions are no-longer valid and the distinction is no longer helpful. Some ethno-linguistic communities engage in self language documentation, recording their own events and language without the participation of outsiders; some researchers may not be affiliated with a university.

18. At the moment, it is unclear if the use of *Deed of Gift* or a similar instrument which transfers the ownership of copyrights and materials to an archive would benefit or hinder OA publications. If the archive is owner of the rights and is also primarily responsible for dissemination, as possessor they could simply restrict access or remove listings of the content in indexes. Presumably though a Deed of Gift would give an archive the ability to issue content under new licenses, which may be more lax than previous releases.

19. Moral rights may not be assertable under U.S. Law, but their exact equivalence and protections when asserted under another legal system is unclear. See the copyright office's statements and report from April 2019. https://www.copyright.gov/policy/moralrights

20. The ultimate destiny of artifacts remains with an archive or preservation organisation as they negotiate the tension between their fiduciary commitments to contributors and their plan for sustainability.

21. See the general arguments presented around intellectual property (IP) rights discussed in the context of the World Intellectual Property Organization and the acknowledgement of IP rights around genetic materials (plants, seeds, and their genomes) and the indigenous knowledge that farmers have about those species.

22. Within the U.S. context some groups of indigenous people (tribes) are recognized by the federal government as another government. These groups have a variety of structural organizational options (Atkinson and Nilles 2008). These organizations can hold property, including intellectual property. Even though many tribes in the U.S. are known (at least to outsiders) by the same name as the name of the language they use or have used in the past, this is not the defining trait of the group in legal or linguistic senses. That is, in the legal

sense, the registered entity of the tribe may hold copyrights, but the amorphous group of tribal members who speak a language may not collectively own the copyright as a collective. They may each hold copyright as an individual, but not as a collective. *Language communities* as a term used by linguists suffers from a lack of precise definition (Patrick 1999, Laitin 2000, Patrick 2008). But the simple fact is that outside of perhaps the U.S., Canada, and Australia, many amorphous groups of language users do not have legal structures, let alone governmental structures which can hold copyrights. For instance in Nigeria there are over 500 languages spoken by ethno-linguistic groups, and Papua New Guinea has over 800 ethno-linguistic groups. Some groups have formed community associations or language committees; however, many of these are not legally registered entities. Even where these committees do exist, they often do not represent the entire community. So for an archive or preservation organization to limit materials to "the community", it raises often unanswerable questions regarding who is or is not part of "the community". Even if these communities do have legal standing at one point in time, they must keep that registered entity current to defend the copyright and to accept new content which may come from new kinds of researchers working on initial archival deposits. Languages exist because people do, not because legal structures exist.

23. For a discussion about the interactions of rights and access at one archive in Australia see Anderson (2005), for an European example see Widlok (2013).

*I am grateful to a variety of former colleagues who have shared their experiences with me including: Will Reiman, Mike Cahill, Bill Dick, Paul Kroening, and Alan Connor. All errors are unfortunate, and are my own.

Hugh Paterson III's experience with sound collections includes having worked at an archive that specifically accessions artifacts which demonstrate language use. In these types of collections, the artifact creation process is generally structured in such a way as to exemplify some language or cultural component, rather than created for entertainment purposes – although sometimes the line between these two perspectives is not well defined, and some overlap does occur. In addition to his work as an archivist, he has also worked as a sound engineer and linguist recording the soundscape produced by speakers of under-documented languages. Under-documented languages is a term used within the domain of linguistics to refer to languages which generally do not have a broad academic description covering their grammar, lexical inventory, speech styles, genres of oral literature, etc. Often these languages have fewer than 100,000 users.