LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 17 (2023), pp. 278-296
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/74738
Revised Version Received: 9 Jul 2023

Assessing access level as a quality criterion

Kristian Roncero
Max Planck Institute for Evolutionary Anthropology
University of Sheffield

This paper discusses levels of access in language archives and their implications for assessment. In the absence of well-established criteria, part of the evaluation of language archives is often based on accessibility; roughly, the more "unrestricted" or "open access" content, the better the archive. In this paper, I argue that whilst OPEN ACCESS may be indicative of good scientific practices, such criteria cannot be extrapolated to documentary linguistics. I start by explaining the differences between those concepts. Then I discuss the dialectic tension in which language depositors frequently find themselves. On the one hand, archives and grant agencies ask them to archive as much as possible with unrestricted access while also gathering a corpus as "naturalistic" and diverse as possible in terms of content, genres, speakers, and so forth. On the other hand, depositors are expected to behave ethically, which involves restricting access to certain materials. I move on to lesser-discussed cases illustrating how ethical decisions concerning accessibility are strongly bound to the needs and situations of the specific communities at a specific time. I finish by proposing new evaluation criteria, which do not penalise depositors for behaving ethically in situations out of their control, and a summary of the discussion.

1. Introduction[1] Peer-reviewing language archives is argued to be the way to tackle the vulnerability of documentary linguists, in terms of academic recognition and career growth opportunities (Haspelmath & Michaelis 2014; Woodbury 2014; Thieberger et al. 2016; Berez-Kroeker et al. 2018; Linguistic Society of America [LSA] 2018; Riesberg 2018; Sullivant 2020). The LSA (2018) issued a statement on evaluation of documentary materials, highlighting their contribution to the science, but also briefly introducing the problems for producing peer-reviewed outcomes for the depositors/researchers. Besides the need for external evaluations (with the incentive for the reviewers of these getting published like any other book review), there are no clear assessment guidelines.

Thus, even though evaluation criteria for linguistic archives are still under development, there are some intuitions that reviewers have been following.[2] However,

[1] Thanks to ELDP grant IPF-0189 for providing the funds to cover my position. Equally thanks to MPI-EVA for providing an extension during the pandemic. A special thanks to all the community members who have worked in the projects I have been involved in, for their kindness and patience with me, particularly to Viktar Alekseevyč and Maryja Handreyŭna. Many thanks to Diana Forker for her support through this project, and thanks to Tihomir Rangelow for reviewing an earlier draft of this paper.

[2] See, for example, the special issue on language archive evaluation criteria from the *Journal of Open Humanities* (forthcoming).

E-ISSN 1934-5275

a careful examination of these criteria is needed, in order not to cause more harm than good. In this case, there is a risk of falling into a faulty comparison between the principles of Open Access – as a sign of good scientific practice – and unrestricted (or open) access to documentary materials in an archive.[3]

Before proceeding further, I should specify the type of evaluation I am referring to. Woodbury (2014) and Woodbury et al. (2015) distinguish three types of reviews for language archives.[4] The first type would be a self-review – an introduction to the archive, like Sophie Salffner's (2015) guide to her Ikaan materials. The second type of review would be an internal review by a tenure or evaluation committee. And the third type, the one in question here, would be an evaluation, which is written (by request of the archive) by an external person who is ideally familiar enough with the area of the represented community, and which later gets published in a journal.

In this paper, I want to follow the discussion opened by Nathan (2014) and Meakins et al. (2018: 91–92) on the dilemma of "open access" or "unrestricted access" to linguistic archive materials. Restricted-access files are often not considered during review and can negatively affect the evaluation. For example, Thieberger et al. (2016: 13) say that "only materials of a corpus that are available to other users can be considered in the review," which means, besides overlooking perhaps useful contents, the depositor may face some uncomfortable criticism from the reviewer for not having more publicly accessible data. Hence, in this paper, I want to discuss the consequences of penalising depositors for restricting access to some of the contents. In §1.1, I compare Open Access in scientific publications to unrestricted access in documentary materials. Since both are very different in nature, the decisions concerning these and their ethical implications will also be very different. I then move on to explain in §1.2 the conflicting situation of language depositors given the plethora of guidelines and stakeholders they need to please, whilst struggling to get recognition for their work for professional growth. In §2, I discuss some situations in the field that are out of the control of the researcher, in which restricting access is the most ethical solution. I want to expand beyond well-known cases, which are often related to copyright, land claims, or the existence of unexploited natural resources, and instead focus on more subtle situations. In §3, I propose new guidelines for archive evaluation that consider the previous discussion around ethics and accessibility. Finally, I close with a summary and extracted conclusions in §4.

1.1 Terminological issues: Open Access vs. unrestricted-access documentary data

With the humanities moving towards models of Open Science, such as the FAIR (Wilkinson et al. 2016) and the TRUST principles (Lin et al. 2020), which focus on data transparency and reproducibility, it is understandable that such principles are permeating throughout language documentation. Nevertheless, assessing the quality of an archive based on its ratio of unrestricted-access materials is a fallacy that is

[3] Hereafter, I use "Open Access" with capital letters to refer to the movement and "open access," "public access" (Seyfeddinipur et al. 2019), or preferably "unrestricted access" to refer to the levels of access in the archives.

[4] Many thanks to the reviewers for pointing this to me.

easy to incur, as the concepts of Open Access in science and unrestricted access to linguistic archive materials differ.

First, it is important to make the distinction between the Open Access movement and the deposit of field materials into an archive with unrestricted access. Suber (2012: 8) summarises the aim of Open Access along the following lines: "The basic idea of OA is simple: Make research literature available online without price barriers and without most permission barriers."[5] Displaying language documentation materials unrestrictedly seeks to contribute to language description, data replicability, and science accessibility to community members. However, Riesberg (2018) argues that Himmelmann's (1998) emblematic paper has been often misinterpreted concerning the display of language documentation materials being unrestricted:

> Turning to the question of replicability, this involves making primary data accessible. […] Needless to say, this is not achieved by simply uploading an audio or video file to a digital archive, or handing over a collection of cassettes or external hard drives to physical one. The falsifiability of descriptive statements was not one of the primary concerns in Himmelmann 1998 […] (Riesberg 2018: 154)

Nathan (2014) argues that whilst the Open Access movement is a push against publishers withholding publicly funded research data behind paywalls (and retaining their copyright), linguistic fieldwork data "are neither scholarly research nor completely reducible to the data that informs that research." In other words, the movement towards Open Access is in part a fight against the unfair practices of big publishing companies and corporations. Conversely, the parties that are negatively affected by the unrestricted access of linguistic recordings are, primarily, the members of the speech community and, to a much lesser extent, the collector/researcher. The CARE principles (Global Indigenous Data Alliance [GIDA] 2018) open with the following statement: "The current movement toward open data and open science does not fully engage with Indigenous Peoples' rights and interests." Nathan (2014) goes further to say that pushing the Open Access agenda "to endangered languages' recordings is a new form of pseudo-scientific colonialism," though I shall be more moderate in my approach.

On the one hand, it is legitimate to push academics who have carried out taxpayer-funded research to share as much data as possible, including linguists working on endangered languages. In some cases, linguists avoid sharing their data due to a lack of tradition within the discipline:

[5] See a more detailed definition in Suber (2012: 3–27).

> […] we realize that some linguists may be reluctant to share data for personal (as opposed to ethical) reasons, and such an attitude is hardly surprising given that data sharing may not previously been standard practice in the subfields many of us work in. (Berez-Kroeker et al. 2018: 15)

Others are still scared of someone reusing their data without attribution and coming up with a better analysis or finding some inaccuracies in their analysis (based on the data they have arduously collected). In real-world terms, this is very unlikely to happen, and there are certain mechanisms for controlling plagiarism and 'data theft' in every peer-reviewed journal or book series. However, this latter concept is very problematic in itself; as by principle, data belong to the speakers/participants/community instead of the researcher/collector (Woodbury 2003).[6]

1.2 Conflicting requirements, tensions between the codes of conduct Besides field-related difficulties, linguists engaging in language documentation face significant challenges with the data they collect. It is so that linguists must satisfy several codes and interests, which often overrun each other and create a dialectic tension as a result. I shall list four of the main messages or expectations for documentary linguists:

a) "Be transparent" As mentioned above, even though reproducibility of statements was not part of Himmelmann's (1998) initial idea, there is a tendency to use field materials (including notes) for these purposes. In any case, the idea behind documentary linguistics is related to the often-invoked image of the linguist who decides to open the archive and study the extinct language in a hundred years (e.g., Bird et al. 2014: 2; Ciucci 2021: 168). Furthermore, the community must be able to see how their knowledge and materials are being handled (GIDA 2018: Principle of Authority to Control).

> The ongoing efforts to document minority languages allow historical linguistics to create, *mutatis mutandis*, a comparable wealth of knowledge for indigenous languages, so that in the medium to long term anyone interested in them can have access to this kind of information. (Ciucci 2021: 168)

Data are also expected to be as 'naturalistic' as possible – that is, representing as faithfully as possible the way the community speaks, which has often been misunderstood as capturing the 'purest' form, or the least-affected by language interference

[6] The CARE (GIDA 2018) Principle of Authority to Control adds the following: "Indigenous Peoples' rights and interests in Indigenous data must be recognised and their authority to control such data be empower […] Indigenous Peoples have the right to develop cultural governance protocols for Indigenous data and be active leaders in the stewardship of, and access to, Indigenous data especially in the context of Indigenous Knowledge."

(see more on this in Szmrecsanyi 2014). In any case, researchers should strive for capturing spontaneous conversations and unprepared texts. Editing should be limited to the very minimum possible. Researchers should also master data-gathering techniques to mitigate as much as possible the so-called observer's paradox.

b) "Be diverse" Documentary linguists/depositors are required to collect not only a plethora of texts, but also these should be as diverse as possible in terms of content, topics, genres, speaker's sociological profile, and so forth (e.g., Berge 2010; Thieberger et al. 2016; Riesberg 2018):

> The content of the corpus should ideally represent a range of speakers across age, gender and other relevant social variables, and a good range of text types. (Thieberger et al. 2016: 16)

c) "Be ethical" Linguists are expected to adhere to the highest ethical standards in research (e.g., Himmelmann 1998; United Nations, General Assembly 2007; GIDA 2018). In our subfield, this means holding human dignity and the well-being of the speech community in the highest esteem. In practical terms, this translates into not recording anything without making the participants involved aware of it and not releasing certain materials or doing so under restricted access. For example:

> […] the compiler of the documentation has to take care that no data are included that may be harmful to an individual or upset the speech community (bad-mouthing, gossip, etc.), even if this possibility is not foreseen by the contributors themselves. (Himmelmann 1998: 172)

Nevertheless, striving for the already mentioned diversity and naturalness in documentary linguistics is ethically problematic. Indeed, as Musgrave & Thieberger (2006: 3) put it, "[a]ll of these recommendations lead to an increase in the intrusiveness of the linguist's practice […]."

d) "Mind the community" The needs and desires of the community should prevail over our "scientific" interests. Researchers must be ready to admit that they can often differ from each other, as Rice (2011: 202) says, "[…] just because the linguist thinks community needs are being met, this does not mean that is the case." Thus, community members must be free all the time to withdraw from the project and ask to restrict access or delete any of their data. Moreover, data ownership must be respected; that is, "data should not be disseminated to those its owners or producers do not want to have or use of it" (Woodbury 2003: 47).[7]

Thankfully, over the past decades there have been many advancements in data protection and ownership for communities in North America, Australia, and New

[7] See also Article 31 of the *United Nations Declaration on the Rights of Indigenous Peoples*, or *UNDRIP* (United Nations, General Assembly 2007).

Zealand (e.g., Christen 2009). See, for example, the aforementioned CARE data-sharing principles (GIDA 2018: Principle C3): "Any value created from Indigenous data should benefit Indigenous communities in an equitable manner and contribute to Indigenous aspirations for wellbeing."

However, researchers working in areas other than these often have still little real accountability to their research institutes and the communities they work with. As a result, researchers have often too much power to do whatever they please with the collected data (and even collect it in unethical ways), which means a lot of the working ethics becomes a matter of conscience and good faith (Wilkins 1992: 171).[8] Moreover, navigating the interface of some language archives can be difficult for the represented communities in rural areas (as they are often only available in English, require a stable broadband connection, and have complex registration steps). As a result, the depositor may face minimal direct consequences, other than betraying the trust of the community. Yet, restricting access to recordings affects the researcher negatively. The work of the researcher is already considered of lesser value than other traditional publication means for career-building purposes (Haspelmath & Michaelis 2014; Berez-Kroeker et al. 2018; LSA 2018). By restricting the access to some of the content, the researcher is not only attracting a potentially worse evaluation of their archive (Thieberger et al. 2016; DELAMAN 2018), but they are also jeopardising opportunities to get more citations on their work[9] – a work which they have often invested many hours recording, editing, transcribing, and glossing.

> Archived data should be made "as open as possible, and as closed as necessary" (European Commission 2016), with consideration of ethical and legal exceptions; though restricted data is citable (and should be cited by those who are permitted to access it), only unrestricted data can be reused and cited […]. (Berez-Kroeker et al. 2018: 12)

Understandably, funders and archive owners are often not happy about having materials with restricted access. The hosting of materials can be expensive, so the curators or owners of the archives want to make the most of their investment – that is, materials under restricted access are considered virtually inaccessible, and hence, "a loss." The recommendation could be to avoid recording certain people, topics, or texts if the researcher knows they will probably only be available under restricted access. However, in some language communities, this would reduce the size of the archive

[8] See, for example, the comment on the first article of the Code of Ethics of the Australian Anthropological Society (AAS): "The objectives of the Code are to provide the consensus view of ethical practice among members […] The Australian Anthropological Society **cannot investigate or adjudicate allegations about unethical behaviour**" (as cited in Musgrave & Thieberger 2006: 3, my emphasis).

[9] According to Woodbury (2014: 32), "[T]he researcher will be subjected to judgement from someone who will never adequately understand the research experience and may make judgements that could be uninformed or inadequately informed formed ones… [t]he review could end up superseding the worth of the collection." In the absence of solidly established evaluation criteria, one of the reviewing 'intuitions' is that "the more publicly accessible material, the better."

considerably, besides compromising the diversity of speakers and genres linguists are asked to collect. Moreover, the community of researchers potentially interested in accessing restricted materials of an endangered language is often very small. Therefore, it is often very easy to control who could be a reliable person to grant access to: either because researchers know each other or because they can be easily tracked using their academic email account and department.

2. Ethical issues Himmelmann (1998: 172), considered a cornerstone of documentary linguistics, already predicted that language documentation, as it is generally conceived, would not be possible in some communities, primarily for ethical reasons.

Ethics is a complex issue that is being discussed in the literature. However, the literature is represented mostly by Native populations of rich Western countries. These communities have made progress raising awareness of their rights and mechanisms to protect these.

> Just as there is no single model for community-based research, ethical standards for community engagement are not universal […] most analyses of endangered language fieldwork situations come from North America. (Holton 2009: 161)

In the following section, I describe some well-known, as well as some hardly discussed issues, which make linguists and depositors consider restricting access to materials as the most ethical and safest solution. As I explain in the archive evaluation criteria in §3, these decisions should bring more respect to the depositor, for being aware of the ethical issues surrounding the community they are working with, rather than penalisation.

The personal examples presented here are based on fieldwork experiences on West Polesian (East Slavic, Belarus), Chamalal (Andic, Nakh-Dagestanian, Russian Caucuses), and, to a lesser extent, remote fieldwork with diaspora speakers of Fa d'Ambö (Portuguese-based creole, Equatorial Guinea).

2.1 Sacred knowledge and traditional knowledge The sacrality of some traditional knowledge has been discussed in the literature for many years (e.g., Himmelmann 1998: 173–176; Battiste & Youngblood Henderson 2000: 132, 140–142; Bowern 2008: 160–161). Hence, I shall only make a few remarks. In some communities, their language, or certain forms of language (e.g., for rituals), or more broadly, some aspects of their traditional knowledge are considered sacred. For this reason, it is important to have a dialogue with the community members about what can be recorded and, out of this, what can be shared publicly (i.e., be unrestricted).

> Ethical research systems and practices should enable Indigenous nations, peoples, and communities to exercise control over information related to their knowledge and heritage and to themselves. […] To act otherwise is to repeat that familiar pattern of decisions being made for Indigenous people by those who presume to know what is best for them. (Rice 2011: 199)

Concerning traditional (nonsacred) knowledge, it is important to remember that it is not uncommon for certain industries (e.g., design, pharmaceutical) to appropriate part of the heritage of more vulnerable people groups, who may not be able to defend themselves and claim their share of benefits. Releasing information as "open access" may not be the most ethical solution in such cases, especially if the community refuses to do so.[10]

2.2 Respecting speakers' choices In some cultures, people are frequently keen on being recorded and photographed, but this is not the case for many other cultures. I have worked with people who, on numerous occasions, have agreed to participate in the project provided that the content would not be posted on larger (and more accessible) platforms (e.g., YouTube) or as an unrestricted-access resource in an archive, or that their image would not be recorded (only voice).

One of the key principles of language documentation is that speakers must have a say in the process; indeed, this *should* be a collaborative process in as much as possible. Rice (2011: 190) refers to the consensus on community-based research along these lines: "community members and researchers equitably share control of the research agenda through active and reciprocal involvement in the research design, implementation, and dissemination."

Thus, speakers *must always* have the right to ask to have access to their recordings restricted, and it is the obligation of the researcher or depositor not to share these with anyone other than the people for whom the data were intended (Woodbury 2003).

> All the contributors will have a say in what can be done and what cannot be done with their contributions. […] I presume without further discussion that the interests and rights of contributors and the speech community should take precedence over scientific interests. (Himmelmann 1998: 172)

The reasons behind speakers opting not to have their recordings publicly available (or only consenting to audio recording) have been very diverse: from body-image issues, to shame of speaking their language in public, to real safety issues (e.g., women being afraid of being seen by certain men), or simply mistrust (e.g., people who lived during Stalin's period who are reminded of the interrogatories and false accusations). Certainly, linguists can spend more time giving details about the project

[10] For a more detailed account of copyright issues in language documentation, see Newman (2012).

and encouraging speakers to reconsider opening access. Even more, this can be an opportunity for participants to receive affirming messages. Nevertheless, linguists should not be expected to solve deep-rooted issues in society or people's lives, and in any case, participants must have the ultimate say.

2.3 Threats to the community Many Indigenous communities live in economically disadvantaged settings and are sometimes ignored or not well tolerated by their rulers. Thus, some community members may feel resentful towards the authorities (often belonging to a different ethnic group). Yet, even though at some point a speaker may share some of their political opinions or denounce corruption, it is the responsibility of the researcher to decide whether it is safe (for the individual or even for the entire community) to display such comments publicly. In fact, if authorities perceive that the archive could be a platform for sharing politically dissident content, they may also penalise the researcher and the archive.

In addition, even though it is also a well-known situation in linguistics and anthropology (e.g., Battiste & Youngblood Henderson 2000), I shall make a brief mention of the issues regarding unexploited resources. Sometimes, speakers may share information about unexploited natural resources, which were unknown to companies. If these companies enter the community, the consequences are almost always terrible for the inhabitants. Thus, if there is the smallest doubt about this, none of the materials should be displayed without access restrictions.

2.4 Asymmetry of locations: Evaluating accessibility based on the area Decisions concerning safeguarding the speakers and the researchers themselves will inevitably look different for every community. Holton (2009: 173) points out that "the pursuit of universal interpretations of ethical guidelines and standards may prove elusive, or in the worst case even harmful." Hence, we cannot force the same interpretation of ethical guidelines for people working in different cultures.

> As Wilkins remarks, "the social, cultural, political, physical, and historical contexts within which linguists do fieldwork are probably more remarkable for their differences than their similarities" (Wilkins 1992:189). Rice is careful to qualify her conclusion with the caveat: "[T]here is not a 'one-size-fits-all' model, as different situations call for different types of interactions" (2006:150). (Holton 2009: 162)

However, this also means that even working in adjacent communities risks making a faulty comparison of the decisions. As logical as it may seem for the reviewers at first glance, language archive assessments should not evaluate the quality of an archive in terms of the ratio of unrestricted-access contents of another archive from the same region. For example, in Dagestan, the situation between different ethnic groups or areas changes drastically within a few kilometres. Southern Dagestan is more liberal with a fair degree of secularisation, whilst Northern Dagestan (especially the mountainous area) is moving towards religious fundamentalism. Thus, re-

cordings covering apparently "innocent" topics, such as traditional music and dance, are well tolerated and even appreciated in Southern Dagestan, but they must have restricted access in mountainous (central-northern) Dagestan.[11] In the same vein, women in the south would be keener on being recorded and participating in the research. Conversely, in the most conservative villages, I had to edit many videos to remove frames in which certain women (who were not allowed to be on the camera) appear (Roncero 2019–2021).

It is also important to remark that the criteria for restricting access can drastically change over time for the same community. As Conathan (2011: 238) puts it, "[o]ver time, the importance of records may change and records may be put to anticipated uses." For example, based on the experiences of other colleagues in the same area of Dagestan, ten to fifteen years ago it was possible to record a wide variety of texts. However, as said, over the past decade there has been a strong move towards religious fundamentalism. Topics such as the aforementioned and recordings of dances or traditional rituals (e.g., the Feast of the First Furrow) would have been acceptable, and the community would have been eager even to display these openly less than seven years ago, whereas nowadays any person sharing such content would be more exposed. In an increasingly connected world, which has been deeply affected by conspiracy theories during the pandemic, we should expect to see more communities in the world that can rapidly switch towards one side or the other. Thus, not only is comparing archives of the same area according to their ratio of open-access materials unhelpful and unfair, but also, given the possibilities of sudden social changes, even two archives representing the same language/group collected within twenty years of each other could vary radically in terms of what is considered safe and acceptable content.

2.5 Sensitive comments of personal character Any materials containing embarrassing comments (about participants or other people) or comments potentially compromising someone's dignity or safety should never be publicly displayed, even if the participant may have agreed to do so. Even though this is understood to be a well-established principle (e.g., Himmelmann 1998: 172), in practice, it is harder to abide by, and the aforementioned penalisations to depositors do not help. It is difficult to imagine a corpus of twenty hours of recordings in which there is not a single sensitive comment (e.g., talking poorly about someone in the village, making paternity claims, or sharing how much they earn).[12] Especially, the more the researcher strives for 'naturalistic' (i.e., data that attempt to mitigate the observer's paradox)

[11] Note that there have been recent reports of online abuse and threats in the region directed towards people 'caught' (on social media) at concerts, as certain people consider music to be a sin. The following articles provide a couple of examples: https://eadaily.com/ru/news/2020/10/30/zaversheno-sledstvie-o-podgotovke-terakta-na-koncerte-kirkorova-v-dagestane (accessed 2021-08-25) and https://md-gazeta.ru/news/52944 (accessed 2021-08-25).

[12] I admit that this point is very controversial and that some of the comments that would be flagged as problematic by white Western scholars (e.g., paternity) may not be problematic in certain communities (see, e.g., McKinnon 2005: 72–119).

and diverse texts, the more likely it is for the speakers to start to share intimate information (Musgrave & Thieberger 2006).

In my experience, there is a risk with certain speakers who become too relaxed and start sharing deeply personal information despite being recorded. Other field linguists have reported similar experiences (e.g., Margetts & Margetts 2012: 15; Meakins et al. 2018: 91). In my case, some of the interviewed women had been victims of domestic violence, and perhaps, I was one of the few people with whom they could talk about it. However, it is our duty to avoid recording such interactions and handle the information entrusted to us with care:

> While it is good when people relax and forget the machine, this increases the chance that they will say or do things they may not want to have recorded, transcribed or played in public. It is the responsibility of the researcher to be sensitive to such situations, and if this is called for, to delete parts of the recording. (Margetts & Margetts 2012: 15)

From an outsider scholar's perspective (and thus, probably, one of the potential reviewers of the archive), displaying such information (unrestricted) may not have much impact because it is unlikely that they know the people involved. However, making such comments unrestricted can be very damaging for members of smaller communities. Therefore, there should be a further debate on access restriction that takes the community's perspective more into account. Furthermore, in an increasingly digitalised world, we should not assume that community members will never be able to access the data (for reasons such as not having internet) and that thus we should not take privacy and safety so seriously.

2.6 Political correctness This point is closely related to the previous one. In some remote areas, people are less concerned about "social advancements," particularly in the Western world. Hence, it is not uncommon to hear or record texts containing "not very politically correct" comments or terminology (e.g., concerning people from other ethnicities, noting differences in ability, or opposing the use of vaccines). In these communities, these comments may be perceived as normal (indeed, they have been the norm in most Western countries until very recently), and it is unfair to expect them to be up-to-date with the latest terminology. Thus, it is a challenge to decide what and how their comments should be displayed in the corpus. For example, should a song or a story not be allowed to be openly displayed in the corpus because it is sexist? If so, then most linguistic archives would have to restrict a significant part of their content. In my case, should I allow to display a story about the deportation of Jewish people in a village (which may be the only evidence that will ever be recorded about them), because at some points the speaker did not use the most commonly accepted term?[13] What shall we do with racist, ableist, sexist jokes in the

[13] Note that the term *žyd* 'Jew' is considered offensive in contemporary Russian. The term *ebrej* '(lit.) Hebrew' is preferred instead. However, in rural areas of Western Polesie, older people use the first one, probably because of contact with Polish, in which it is a neutral term.

corpus if these are perceived as "unproblematic"' by the speech community?[14] This is certainly a delicate issue, which has to be examined on a case-by-case basis, and there is a risk of falling into paternalism. However, as a general criterion, I believe that texts deliberately attempting against someone's (or a group of people's) dignity, such as the above-mentioned distasteful jokes, should not be made publicly accessible. Also note that if the archive materials are shared on social media, these may get blocked for the same reasons. That being said, there is a value in recording and documenting every form of language as it may be useful for some. In the same vein, someone working on discourse analysis (e.g., hate speech) would have to examine similar contents in other languages.

Finally, I would like to mention the paradoxes of our "Western morality filter," especially regarding bodily issues (not to even mention nudity). For example, in some cultures, picking your nose, scratching certain body parts in public, or producing certain sounds is deemed unproblematic. When a speaker has done something like this on a recording, it has been challenging to decide what to do. The CARE data principles (GIDA 2018: Principle E2) say that "[e]thical data are data that do not stigmatize or portray Indigenous Peoples, cultures, or knowledges in terms of deficit." On the one hand, if such things are acceptable in the community, we should not impose our moral agenda on them and cause more harm. On the other hand, people in urban areas of the same region may be more Westernised and despise such things. As a result, releasing a video containing a speaker in a "compromising situation" could be considered as attempting against their dignity and help to perpetuate the stigmatisation of the community and perception that they are "less civilised."

3. Evaluation criteria in the light of access control and ethics Concerning tangible steps for archive evaluations, I propose the following: Simply restricting access to the files without an explanation can be problematic, and making the justification for restricting public access can still be dangerous for the participants. For example, some recordings in my archive had to be restricted because they contained texts or comments about music, which, as I have mentioned, is a sensitive topic in the area. However, if the codes of the participants are displayed and the topics and/or keywords shown, the participants can still be tracked (because their codes are public in other files). For this reason, to achieve more transparency but remain ethical, archives should provide a comment box only for internal use (and that is visible to the evaluators). In this box, the codes of the participants and a detailed explanation for restricting access would be available. Archive reviewers should take into consideration the details of this section and be granted full access to restricted materials for evaluation purposes. This comprises more technical difficulties for the design of the archives, but it is something to aim for.

Now, building upon the CARE (GIDA 2018) and TRUST (Lin et al. 2020) principles and the criteria given by DELAMAN (2018) and Musgrave & Thieberger (2006), I would like to propose a more elaborate list of criteria to evaluate language

[14] It is important to note that Article 22 of the *UNDRIP* (United Nations, General Assembly 2007) is also mindful of these.

documentation archives concerning good practices for accessibility and openness. These assessment criteria are not an end in themselves, but rather a means to guide the reviewer and, most importantly, start a conversation on the practical implications of research ethics and data management in less frequently discussed field sites.

The first three general assessment criteria I will propose are usually monitored by the archivist, but it is still the responsibility of the depositor to make sure that everything is in order. First, there should be informed consent available from all the participants (note that depending on local data protection regulations, some of the data in these may have to remain restricted). Second, all files should be in "standard formats" (Musgrave & Thieberger 2006: 5) – that is, "non-proprietary formats" (Thieberger et al. 2016: 13) or "open formats" (e.g., Boas 2006; Good 2011) – which are also accessible for the community. Third, there should be rich metadata, which allow relevant documents to be easily found but are not too technical (or inaccessible for the community, if they speak the metalanguage of the archive). Yet, if needed, the metadata of potentially compromising files should allow the contributors' information to be anonymised for their safety.[15]

The second set of assessment criteria has to do with academic outputs, provided there is community consent to display these (cf. Wilkinson 1992) and there are no embargoes from the publisher. Besides 'pure documentary recordings' (i.e., free texts), recordings from elicitation sessions and a summary or notes from which the data are derived should be provided (e.g., Roncero 2020). Recordings containing paradigms, grammar questionnaires, and translations are frequently less ethically problematic than free texts (Meakins et al. 2018: 91), and speakers are often more open about audio-only recordings being displayed unrestricted. Unless there are ethical issues, materials should be displayed unrestricted for data transparency and reproducibility. This can be challenging as it also exposes some of the flaws and misunderstandings of the researcher, but it helps to understand where the data came from and track some of the potential inaccuracies.

The third set of assessment criteria is related to the use of access protocols. A good archive will most likely use different levels; that is, some recordings are not openly available, and when it is possible to do it privately or without compromising the safety and dignity of the participants, a justification is given. The researcher is sensitive towards the represented culture and shows the capability to judge which materials are appropriate for public display.

Note that my discussion is primarily between two extremes: UNRESTRICTED vs. RESTRICTED. I am aware that some archives include the possibility of only displaying

[15] "TRUSTworthy repositories should enable their community to find, explore, and understand their data holdings with regard to potential (re)use. Repositories should encourage users to fully describe data at the time of deposition and facilitate feedback on any issues with the data (e.g. quality or fitness for use) that may become apparent after the data have been made available" (Lin et al. 2020). Alternatively, the better-known CARE principles state the following: "[d]ata governance should take into account the potential future use and future harm based on ethical frameworks grounded in the values and principles of the relevant Indigenous community. Metadata should acknowledge the provenance and purpose and any limitations or obligations in secondary use inclusive of issues of consent" (GIDA 2018: Principle E3).

the content to registered users. However, given the frequently complex registration steps for community members (e.g., the interface is often only in English and they require an email account),[16] in practical terms, materials displayed as such are not accessible for many community members, yet this does not protect them from some of the issues discussed in §2. I am also aware of more complex and community-sensitive customised access protocol forms (e.g., only for initiated or only for women), such as the Mukurtu project (Christen 2009). Yet given the aforementioned reasons, besides the costs of creating a tailor-made archive (which most documentation projects cannot afford) and the need for community members who are fluent enough in the metalanguage of the archive, customised access controls do not apply to people depositing in larger archives with a set structure. In any case, if such access levels exist, they should be included in the evaluation (see criteria below in Table 1).
I have summarised these criteria in a checkbox in Table 1.

Table 1. Evaluation criteria[17]

Criterion	**Y/N**	**Why?/Remarks**
General (related to archive)		
1. There is recorded or written informed consent from the participants.		
2. Are the archived materials in standard/ nonproprietary formats?		
3. The metadata are rich enough to easily find the relevant queries. Provide examples of specific search terms.		
3.1 Search terms are not too obscure or technical for the community (if they speak the metalanguage of the archive).		
3.2 There are no examples of metadata of restricted recordings that could expose somebody.		
Data transparency/replicability		
4.1 Are there any derived published academic materials openly available?		Is there an embargo from the publisher?
4.2 Are there any derived unpublished academic materials openly available?		Is there community consent to display these?
5. Are the field notes and purely grammar elicitation recordings available?		

[16] See also Thieberger et al. (2016: 10).

[17] Many thanks to the reviewers for providing such useful feedback to make this more community-driven.

Access control		
6. Does the researcher/depositor use different levels of access?		
6.1 What are the access levels used? (Judge according to the technical possibilities offered by the repository.)		
6.1.1 Are the access levels standard for the archive or customised for cultural needs (e.g., only for women)?		Does the general platform offer such a possibility? What work was involved to set up these standards with the community and archive? Is there any evidence?
6.1.1.1 Does the community curate the forms of access?		Is there an (approved) member who is competent in the metalanguage?
7. Does the community curate the levels of access?		
7.1 If so, has the researcher set up a contact person/organisation/process in between the community and the archive?		
7.1.1 If not, does the researcher provide justification for the levels and any customised access?		
8.1 Is there evidence of culturally sensitive topics or social biases that is not restricted?		
8.1.1 Does the researcher provide adequate information about these?		

4. Summary and conclusions In sum, Open Access, as a movement against publicly funded research outcomes being held behind a paywall, is not equivalent to providing unrestricted access to language documentation materials. The main beneficiaries of the former are scientists, whilst the latter, unless properly managed, can cause more harm than good to the participants/community and, to a lesser extent, the linguist. Language archive contributors already have a difficult time getting the same recognition for their work as their fellow linguists. Restricting access to any potentially sensitive contents should not translate into an additional penalisation for the depositor, who is already missing out on an opportunity to get more views and citations of the materials they have worked on for so long. This only encourages unethical practices, such as the researcher disregarding possible consequences for the community before sharing content. As explained, the reasons behind limiting the

access to materials can be many, and these are out of the researcher's control in the vast majority of cases.

In any case, the ethical principles for handling language documentation materials should remain the same: "erring on the side of caution" by not making the material openly available (Macri & Sarmento 2010). Thus, archive collections should not be evaluated based on the ratio of unrestricted materials (i.e., the more, the better), nor is it fair to judge them in comparison to other collections of the area on this parameter. In fact, a good, diverse, and ethically collected archive will most likely contain some recordings that are not good for unrestricted display. The number one priority should remain the community and the safety and dignity of the speakers. Hence, the researcher's determination to adhere to ethical standards should be considered favourably rather than negatively.

References

Battiste, Marie & James Youngblood (Sa'ke'j) Henderson. 2000. *Protecting Indigenous knowledge and heritage: A global challenge*. Saskatoon: Purich Publishing.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. doi:10.1515/ling-2017-0032

Berge, Anna. 2010. Adequacy in documentation. In Grenoble, Lenore A. & N. Louanna Furbee (eds.), *Language documentation: Practices and values,* 51–66. Amsterdam: John Benjamins Publishing Company.

Bird, Steven, Florian R. Hanke, Oliver Adams, & Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In Good, Jeff, Julia Hirschberg, & Owen Rambow (eds.), *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Baltimore, 26 June, 1–5. Baltimore: Association for Computation Linguistics. doi:10.3115/v1/W14-2201

Boas, Hans Christian. 2006. From the field to the web: Implementing best-practice recommendations in documentary linguistics. *Language Resources and Evaluation* 40. 153–174. doi:10.1007/s10579-006-9012-6

Bowern, Claire. 2008. *Linguistic fieldwork: A practical guide*. New York: Palgrave MacMillan.

Christen, Kimberly. 2009. Access and accountability: The ecology of information sharing in the digital age. *Anthropology News* (50)4. 4–5. doi:10.1111/j.1556-3502.2009.50404.x

Ciucci, Luca. 2021. From fieldwork to reconstruction: Historical issues in hotspots of linguistic diversity. *Studia Linguistica* 75(2). 165–174. doi:10.1111/stul.12160

Conathan, Lisa. 2011. Archiving and language documentation. In Austin, Peter K. & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 235–254. Cambridge: Cambridge University Press.

Good, Jeff. 2011. Data and language documentation. In Austin, Peter K. & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 212–234. Cambridge: Cambridge University Press.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Holton, Gary. 2009. Relatively ethical: A comparison of linguistic research paradigms in Alaska and Indonesia. *Language Documentation & Conservation* 3(2). 161–175. (http://hdl.handle.net/10125/4424)

Macri, Martha & James Sarmento. 2010. Respecting privacy: Ethical and pragmatic considerations. *Language & Communication* 30(3). 192–197. doi:10.1016/j.langcom.2009.11.005

Margetts, Anna & Andrew Margetts. 2012. Audio and video recording techniques for linguistic research. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork,* 13–53. Oxford: Oxford University Press.

McKinnon, Susan. 2005. *Neo-liberal genetics: The myths and moral tales of evolutionary psychology*. Chicago: Prickly Paradigm Press.

Meakins, Felicity, Jennifer Green, & Myfany Turpin. 2018. *Understanding linguistic fieldwork*. London: Routledge.

Musgrave, Simon & Nick Thieberger. 2006. Ethical challenges in documentary linguistics. In Allan, Keith (ed.), *Selected papers from the 2005 Conference of the Australian Linguistic Society*, Melbourne, 28–30 September. (http://www.als.asn.au/proceedings/als2005.html) (Accessed 2021-05-25.)

Newman, Paul. 2012. Copyright and other legal issues. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 430–456. Oxford: Oxford University Press.

Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5. 187–207. (http://hdl.handle.net/10125/4498)

Rice, Keren. 2012. Ethical issues in linguistic fieldwork. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 407–429. Oxford: Oxford University Press.

Riesberg, Sonja. 2018. Reflections on descriptive and documentary adequacy. In McDonnell, Bradley, Andrea L. Berez-Kroeker, & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998* (Language Documentation & Conservation Special Publication 15), 151–156. Honolulu: University of Hawai'i Press. (http://hdl.handle.net/10125/24816)

Roncero, Kristian. 2019–2021. *An audiovisual corpus of Chamalal, a language of Dagestan*. Berlin: Endangered Languages Archive (ELAR). (https://elar.soas.ac.uk/Collection/MPI1234834) (Accessed 2023-01-06.)

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267. (http://hdl.handle.net/10125/24639)

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, et al. 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. 2012. *Open access*. Cambridge, MA: The MIT Press. doi:10.7551/mitpress/9286.001.0001

Sullivant, Ryan. 2020. Archival description for language documentation collections. *Language Documentation & Conservation* 14. 520–578. (http://hdl.handle.net/10125/24949)

Szmrecsanyi, Benedikt. 2014. Methods and objectives in contemporary dialectology. In Seržant, Ilja A. & Björn Wiemer (eds.), *Contemporary approaches to dialectology: The area of North, Northwest Russian and Belarusian dialects*, 81–92. Bergen: University of Bergen.

Thieberger, Nick, Anna Margetts, Stephen Morey, & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36(1). 1–21. doi:10.1080/07268602.2016.1109428

Wilkins, David. 1992. Linguistic research under Aboriginal control: A personal account of fieldwork in central Australia. *Australian Journal of Linguistics* 12(1). 171–200. doi:10.1080/07268609208599475

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

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation and Description* 1. 35–51. doi:10.25894/ldd305

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In Nathan, David & Peter K. Austin (eds.), *Language Documentation and Description* (Special Issue on Language Documentation and Archiving), vol. 12, 19–36. London: SOAS.

Woodbury, Anthony C., Richard P. Meier, Sarah G. Thomason, Meagan Dailey, Ryan Henke, Jaime Perez Gonzalez, Nick Williams, & Susan Smythe Kung. 2015. Reviewing and evaluating archived collections. In *Presentations from the 1st Workshop on Data Citation and Attribution in Linguistics*, Boulder, 18–20 September. (http://hdl.handle.net/10125/36885)

Other Resources

Digital Endangered Languages and Music Archiving Network (DELAMAN). 2018. *Minimal checklist for the preservation of digital language documentation materials*. DELAMAN. (http://hdl.handle.net/10125/55829)

EurAsia Daily. 2020. Zaveršeno sldestvie o podgotovke terakta na koncerte Kirkorova v Dagestane [The investigation into the preparation of a terrorist attack at Kirkorov's concert in Dagestan is completed]. *EurAsia Daily*, October 30. (https://eadaily.com/ru/news/2020/10/30/zaversheno-sledstvie-o-podgotovke-terakta-na-koncerte-kirkorova-v-dagestane) (Accessed 2021-02-13.)

Global Indigenous Data Alliance (GIDA). 2018. *CARE Principles for Indigenous Data Governance*. (https://www.gida-global.org/care) (Accessed 2021-09-25.)

Haspelmath, Martin & Susanne Maria Michaelis. 2014. Annotated corpora of small languages as refereed publications: A vision. *Diversity Linguistics Comment*. (https://dlc.hypotheses.org/691) (Accessed 2021-05-17.)

Lin, Dawei, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, et al. 2020. The TRUST Principles for digital repositories. *Scientific Data* 7. 144. doi:10.1038/s41597-020-0486-7

LSA (Linguistic Society of America). 2018. *Statement on evaluation of language documentation for hiring, tenure, and promotion*. (https://www.linguisticsociety.org/sites/default/files/Evaluation_Lg_Documentation.pdf) (Accessed 2021-06-01.)

Molodež' Dagestana. 2018. *«Toplivo dlja ada». Nurmagomedov i Maxačev osudili zritelej rèp-koncerta v Maxačkale* ["Fuel for Hell." Nurmagomedov and Makhachev condemned the spectators of the rap concert in Makhachkala]. August 15. (https://md-gazeta.ru/news/52944) (Accessed 2021-02-13.)

Nathan, David. 2014. The Access powers: Open access and language documentation. *Endangered Languages and Cultures*, March 10. (https://www.paradisec.org.au/blog/2014/03/the-access-powers-open-access-and-language-documentation-david-nathan/) (Accessed 2021-05-19.)

Roncero, Kristian. 2020. *Fieldwork session planner* (v.2). doi:10.5281/zenodo.3835660

United Nations, General Assembly. 2007. *The United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP)*, A/RES/61/295. (https://documents-dds-ny.un.org/doc/UNDOC/GEN/N06/512/07/PDF/N0651207.pdf) (Last accessed 2023-06-09.)

Kristian Roncero
kristian.roncero@gmail.com
orcid.org/0000-0001-8198-7393