Public access to research data in language documentation: Challenges and possible strategies

Mandana Seyfeddinipur, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L Epps, Vera Ferreira, Ana Vilacy Galucio, et al.

LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 13 (2019), pp. 545–563
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24901
Revised Version Received: 5 Aug 2019

Public access to research data in language documentation: Challenges and possible strategies

Mandana Seyfeddinipur
SOAS University of London, UK

Felix Ameka
Leiden University, The Netherlands

Lissant Bolton
British Museum, UK

Jonathan Blumtritt
University of Cologne, Germany

Brian Carpenter
American Philosophical Society, USA

Hilaria Cruz
University of Kentucky, USA

Sebastian Drude
Clarin, The Netherlands

Patience L. Epps
University of Texas Austin, USA

Vera Ferreira
SOAS University of London, UK

Ana Vilacy Galucio
Museu Paraense Emilio Goeldi, Brazil

Brigit Hellwig
University of Cologne, Germany

Oliver Hinte
University of Cologne, Germany

Gary Holton
University of Hawaii, USA

Dagmar Jung
University of Cologne, Germany

Irmgarda Kasinskaite Buddeberg
UNESCO, France

Manfred Krifka
Leibniz Zentrum Allgemeine Sprachwissenschaft, Germany

Susan Kung
University of Texas Austin, USA

Miyuki Monroig
World Intellectual Property Organization, Geneva

Ayu'nwi Ngwabe Neba
University of Buea, Cameroon

Sebastian Nordhoff
Free University Berlin, Germany

Brigitte Pakendorf
Université de Lyon, France

Kilu von Prince
Leibniz Zentrum Allgemeine Sprachwissenschaft, Germany

Felix Rau
University of Cologne, Germany

Keren Rice
University of Toronto, Canada

Michael Riessler
University of Freiburg, Germany

Vera Szoelloesi Brenig
Volkswagen Stiftung, Germany

Nick Thieberger
Paradisec, University of Melbourne, Australia

Paul Trilsbeek
Max Planck Institute for Psycholinguistics, The Netherlands

Hein van der Voort
Museu Paraense Emilio Goeldi, Brazil

Tony Woodbury
University of Texas Austin, USA

E-ISSN 1934-5275

The Open Access Movement promotes free and unfettered access to research publications and, increasingly, to the primary data which underly those publications. As the field of documentary linguistics seeks to record and preserve culturally and linguistically relevant materials, the question of how openly accessible these materials should be becomes increasingly important. This paper aims to guide researchers and other stakeholders in finding an appropriate balance between accessibility and confidentiality of data, addressing community questions and legal, institutional, and intellectual issues that pose challenges to accessible data.

1. Introduction Over the past two decades Open Access to research publications has become increasingly valued by researchers, funding organizations, and the general public.[1] There is an increasing expectation that the products of publicly funded scientific research should be open to all. More recently this expectation is being extended not only to the products of research but also to the primary data from which those results derive. Providing access to primary data facilitates reproducible research, ensuring scientific accountability for research results while also increasing transparency, efficiency, and collaboration (cf. Berez-Kroeker et al. 2018). Another type of challenge arises from statements such as the Berlin Declaration on Open Access,[2] which affects Open Access publications. The Berlin Declaration requires that "[t]he author(s) and right holder(s) of [Open Access] contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship". While not legally binding, such declarations can conflict with community interests where limitations on access might be important, or where communities are concerned that their materials might be misappropriated and used for commercial purposes.

The issues raised by the Open Access Movement are impacting all areas of linguistics, but they are particularly significant within documentary linguistics, given the focus of this subfield on primary data. This paper discusses issues surrounding public access to data produced by language documentation projects, i.e., projects which create collections of annotated recordings of people speaking about their lives, cultures, and histories. The tensions arising from the nature of the projects are manifold and relate to privacy and copyright issues, among others (cf. Janke 1998; Brown 2003; Thieberger & Musgrave 2007).

Since the emergence of documentary linguistics as a sub-discipline in the late 1990s, recording and preserving culturally relevant materials, natural dialogues, and oral literature have been important for research, documenting and preserving cultural heritage, and providing community members with access to language data. Accessi-

[1]A chronological overview of the Open Access Movement can be found at https://legacy.earlham.edu/ peters/fos/timeline.htm (Accessed 21 May 2019) – a timeline created by Peter Suber (one of the Open Access pioneers), which covers the period up to 2008. Beyond 2008, this timeline was continued in wiki form at the Open Access Directory and can be consulted at http://oad.simmons.edu/oadwiki/Timeline (Accessed 21 May 2019). A visualised timeline is also available at https://symplectic.co.uk/open-access-timeline/ (Accessed 21 May 2019). For a critical reflection on the definition(s) of Open Access and its implications for indigenous knowledge sharing see Christen (2012), and Singer (2014).

[2]https://openaccess.mpg.de/Berlin-Declaration (Accessed 10 April 2018).

bility is fundamental to the field of documentary linguistics; as summarized by Himmelmann (1998:165), "it is simply a feature of a scientific enterprise to make one's primary data accessible to further scrutiny". However, while Open Access might be seen as an ideal from the open research perspective (OECD 2015), *fully* open data are not always possible or desirable from a cultural, ethical, and privacy perspective (cf. Dwyer 2006; Rice 2006; 2011; Austin 2010; van Driem 2016, among others, for detailed discussions on ethical issues in language documentation).[3] This is because language documentation projects typically produce audio and video recordings which may contain personal or politically sensitive content, or material that is culturally inappropriate to share (cf. Brown 2003:229ff; Christen 2012:2875). This content consists of a variety of genres of natural speech, including traditional stories, histories, cultural activities, procedural accounts, conversational interactions between people, and traditional knowledge, as well as gossip, personal stories, and political discussions. We need to be aware of the colonial nature of academic research, as "imperialism and colonialism brought complete disorder to colonized peoples, disconnecting them from their histories, their landscapes, their languages, their social relations and their own ways of thinking, feeling and interacting with the world" (Smith 1999:28). The role of archives in making material available can be seen as both a continuation of neocolonialist methods, and as a postcolonial repatriation, because restricting access to primary records, which academics are often criticized for, is also seen as bad practice.

Following Christen (2012:2883), "knowledge can (and does) die if it is not used. But it also needs to be used and circulated within an articulated ethical system". Because of the nature of the content of the recordings, access to them may be restricted for several reasons. From a community perspective, recordings may be considered sensitive and not appropriate for Open Access because of their personal or political nature or because knowledge is not seen as shareable with non-community members (cf. Christen 2015). Moreover, researchers might fear that data made publicly available before they fully analyze it may be mined by others who will scoop the original researcher.[4]

Responding to these concerns, many digital archives working with endangered language materials and communities have implemented graded access restrictions.[5] In some instances, depositors are able to specify who should have access to their recordings. In addition, most archives require users to agree to an ethical code of conduct prior to accessing materials, or they may restrict use to educational or academic non-commercial purposes. Strictly speaking, these types of restrictions do not constitute Open Access, as they place an additional barrier between the user and the data and may restrict the way the materials are used and repurposed. For the pur-

[3]It should be mentioned at this point that the Open Access Movement is not trying to make everything open regardless of sensitivities and nuances. Even the strongest supporters of Open Access recognize that open access is not appropriate for every situation.

[4]This fear is reflected in the tendency for PhD students to put embargos on data deposited with language archives. This shows, furthermore, that scooping in itself is more a problem of the academic career and less a problem of the reusage of data.

[5]Examples of such archives are those that are members of the Digital Endangered Languages and Musics Archives Network (DELAMAN). http://www.delaman.org/. (Accessed 10 April 2018).

poses of this paper we will refer to this type of access as Public Access. Some archives may place further restrictions on access to some items, such as requiring users to request access to recordings directly from the depositor. This type of access would not be considered public access.

This paper aims to guide researchers and other stakeholders in finding an appropriate balance between accessibility and confidentiality of data, addressing community questions and legal, institutional, and intellectual issues that pose challenges to accessible data. The paper is organized as follows. We first address issues around communities in §2, then turn to legal issues and ownership of data in §3. Following this, we examine institutions and public access, including a discussion of costing models and archives, in §4. We then turn to data types and the access challenges connected to them in §5, and end with a discussion of credit and control in §6. In all cases, we first set out some of the challenges posed by the goal of public access, and then identify strategies as recommendations that might be used to address those challenges.

2. Communities and public access

This section introduces the types of community issues that may arise from public access to language documentation data and examine some strategies that can be used to address these issues.

2.1 Challenges

Communities and researchers are often concerned about certain types of material being made publicly available. This could be because the material is sacred, spiritual, or even secret in content, is intimately connected to communities' traditional knowledge and genetic resources, or because the material is politically sensitive or identifies individuals in ways that are potentially harmful to them. Communities may be suspicious about how publicly accessible material might be used, and how outsiders might profit from the material. A further challenge arises from the question of how to ensure, in regions with little or no internet access, that the concept of worldwide digital sharing can be explained, with all its consequences.

Community perspectives and concerns about notions of authorship and intellectual property rights, and who has the right to determine to whom material can be made available, may also vary (cf. Whimp & Busse 2000). Determining who has the cultural and legal authority to provide consent may be complex where individuals, families, or communities hold rights to specific stories, songs, dances, or other cultural expressions. Community membership and rights to speak for the community may be contested. Legal and cultural rights and authorities may be affected by clan membership, gender, and individual issues, and there may be groups or institutions within the community who compete for authority. There is also regional variation in attitudes to ownership and control of knowledge and language. In some places language is owned and knowledge must be bought, or used only by the knowledge-holders (cf. Wilkins 1992). Researchers or depositors must be aware of these specificities and provide information about such attitudes in their collection metadata, so that archive staff and users are aware of them.

2.2 Strategies The concerns raised above can be addressed through discussions within the language community, and by working together to implement an ethical framework for ownership, intellectual property and access. Informed consent – ensuring that speakers are aware of the potential harm caused by their participation in a language documentation project – can provide a vehicle for addressing some community concerns. It entails that speakers determine ownership and who will have access to materials resulting from the documentation. Whether informed consent is mandatory because of conditions set by a university, a funder, or a community, discussing issues around consent is essential in understanding intellectual property rights and access. See Fluehr-Lobban 1994, Grinevald 2006, and Robinson 2010, among others, for detailed discussions on informed consent.

Ownership of material and questions around access options need to be discussed early, both with individuals and the wider community, with discussion continuing on a regular basis, and these discussions should be situated within an appropriate ethical framework. Questions such as the following can be considered in the process of understanding and dealing with the specificities of ownership of the data collected: Is this story one that anyone in the community has the right to tell? Does this version belong to a particular person, while in some sense it also belongs to a family?

What level of access the speaker or the community wishes to give to materials resulting from documentation is another topic that needs attention. Here are some important questions to be considered when discussing this issue: Who can listen to, view, or read particular materials, and what does it mean if anyone in the world could do this? Can only a family or a family member listen to, view, or read a story? Could people from a neighboring village listen to, view, or read this material? What about someone from a more distant urban area? How about a government official? Just what these categories are will differ from place to place, making some degree of ethnographic understanding necessary. Additionally, to deal with access issues from within the community, one should also ask beforehand how data made available on the internet might be used.

Workshops can be held to discuss these topics. Notions of authorship, ownership, and accessibility, addressing questions such as those given above, can be discussed. Training can be provided for individual speakers, who can then explain the issues to others. Examples from existing archives can be used to highlight what an archive is, how authorship is indicated, and conditions on access. Likewise, researchers can be educated as to community concerns about access.

More formally, consent should be documented in an appropriate form for the individual and the community: Where written consent is not suitable, speakers' agreement can be recorded orally, as can relevant discussions with the wider community. Community sensitivity to material may vary depending on its format – video, audio, or written – and linguists and archives should be aware that community restrictions might in fact apply only to particular components of a given data set.

Any consent obtained should take into account both authorship and access conditions. Individual and community views of consent can change over time, and these issues should be discussed around any recordings that might be viewed as sensitive,

either with respect to authorship or use. Some material may be deemed inappropriate for archiving and may thus be retained by communities or individuals, or else destroyed. Recordings that were not deemed sensitive at one time might come to be viewed as such at a later point in time and vice versa, so these issues must be revisited regularly in order to ensure that community and individual interests are respected and that appropriate access levels are implemented. Therefore, informed consent should include discussion of the level of access (open, or restricted in some way), and this discussion should be included as part of the collection's metadata.

In some (or perhaps many) cases, truly "informed consent" around access may be unachievable, as the concept of worldwide digital sharing, its scope, and the potential for materials to be misused or misinterpreted is not easily explained. The aim is for informed consent to be as informed as possible. It may be appropriate to err on the side of caution and restrict access, at least in the early stages of research.

Further considerations relate to potential uses of the material that may violate community interests and access agreements. For example, there is a risk that ethnobotanical or artistic material drawn from Open Access deposits could be used in ways that fail to recognize community intellectual property rights, and even for commercial gain – in spite of explicit licenses which prohibit such uses. These risks can be at least partially mitigated by archive-based requirements for registering users, tracking downloads to allow better oversight of the use of the content, and providing clear ethical guidelines on legitimate uses of the material. These risks also need to be weighed against the colonial legacy of withholding materials from the people who have a direct interest in them.

3. Legal issues, ownership, and public access Just as communities can challenge Open Access to materials, legal and ownership issues also present challenges. This section introduces some of these challenges.

3.1 Challenges In some jurisdictions research permits are required in order to conduct a language documentation project, and the permits may place explicit restrictions on access to research data. Where permit processes require researchers to guarantee that research outcomes will not be used for non-research related purposes, particularly commercial gain, violations (actual or perceived) may lead to the loss of a permit and to further implications for a researcher's career. Many universities also require that an ethics protocol be approved before research can begin. The research cannot take place without the permission of the appropriate people or institutions (cf. Bowern 2010; O'Meara & Good 2010; Næss & Hovdhaugen 2011; Good 2018).

Legislation regarding research data varies according to jurisdiction. In some countries there is a requirement that research data (particularly data seen as including personal information) be destroyed once the research is complete. If language documentation data is not exempt from this, a justification can be made for its preservation in an archive, which must happen before a researcher collects the data and requires informed consent to do so, as mentioned in §2. In some publicly-funded archives all archived material is required by freedom of information laws to be made openly

available upon request, as is the case, for instance, with recorded information held by public authorities in England, Wales and Northern Ireland, and by UK-wide public authorities based in Scotland.

Different ethical standards and regulations governing access and copyright may have repercussions for collaboration and working across international boundaries. Researchers must observe the local legal frameworks that apply in all countries where they work, conforming to data protection and privacy laws, obeying national copyright regulations and intellectual property rules, and respecting freedom of information laws. Intellectual property rights may apply differently to original recordings and written texts, as opposed to transcriptions, translations, and other annotations.

3.2 Strategies Researchers should be aware of legal issues and requirements in their institutions, resident countries, the countries and communities in which they conduct research, and the countries in which work will be archived. It is also important to keep in mind that where research permits are required these might include restrictions on data use and access. Researchers should also be informed about these requirements well in advance of beginning the research.

Moreover, researchers must also understand the intellectual property implications of documenting traditional knowledge. Traditional knowledge refers to the "knowledge, know-how, skills and practices that are developed, sustained and passed on from generation to generation within a community" (cf. WIPO 2016a). Due to its low level of legal recognition in many countries, traditional knowledge is not easily protected by the current intellectual property system, which "typically grants protection for a limited period to new inventions and original works by individuals or companies" (cf. WIPO 2016a). Intellectual Property law typically vests copyright in language documentation materials with the individuals who made the recordings – i.e., linguists, anthropologists, etc. – rather than the speakers. This means traditional knowledge holders do not have legal ownership over the materials and cannot determine their legal use (see Macmillan 2013 for a discussion about legal protection of tangible and intangible cultural heritage; see also Khan 2018). In this sense, prior informed consent is essential to clearly assign copyright to speakers, negotiate appropriate licensing, and ensure that communities and individuals can exercise rights over the material provided and that these are acknowledged accordingly.

One strategy for avoiding the strongest implications of the Berlin Declaration for Open Access publications and similar documents is to use the Creative Commons Non-Commercial license.[6] This prevents material from being used in textbooks available for sale, in language schools which charge fees, and on websites which run advertisements to finance costs. However, this strategy may also limit reuse where language initiatives rely on such income streams to finance their operations. Perhaps a more effective strategy for avoiding the commercial use of materials is the Creative Commons Share-Alike license, under which all derivative content must again be made

[6](Accessed 10 April 2018).

freely available.[7] This means, for instance, that a movie made using content from the collection must be available under the same open license.

Different archives have different license or "deed of gift" standards. Some require that copyright be assigned or licensed to the archive, while others stipulate that data creators or authors retain copyright. Other archives require the depositor to apply a Creative Commons license to their research publications.

4. Institutions This section examines institutions broadly, including archives. Issues relating to access, data types, and users of archives are addressed below in §5.

4.1 Challenges Public access to research data requires long-term archiving of language data. This in turn requires a long-term commitment by institutions to maintaining and developing technology to sustain archives and avoid data graveyards. This involves costs, and institutions require models to meet those costs over a sustained time period.

Currently, systematic standardized policies concerning data management and accessibility for funders, researchers, and archives are lacking. Such policies would entail creating interfaces and developing the usability of archives, while meeting high standards for deposits, with reports on usage and impact. There is little training available yet in this kind of data management (see Gawne et al. 2017).

Archiving and maintaining archives comes at a cost, and there is a cost to providing high quality presentations and interfaces, but there is also a cost to not doing so (see Thieberger 2014). Digital archives must be maintained and offer new functions, services, and modes of display that make the data as accessible as possible.

4.2 Strategies One strategy for resolving this challenge is funding. If funding were available to support the work institutions need to do, the skills and talent could be found to do it. Institutions involved in archiving (including museums, galleries, archives, libraries, and research centers) need to collaborate to identify common solutions, both in technology and costing, to ensure continuing support. Systematic, standardized policies concerning data management will be of value to funders, researchers, and archives. The following suggestions should be incorporated into the workflows of the institutions dealing with archiving:

- Restricted access must be justified. See §5.2 and §6.
- Data management, curation, archiving, and publishing should be properly budgeted for beyond a project's lifespan.
- Embargo periods for primary researchers should have time limits and should expire unless a longer time period is explicitly sought. See §6.2.2.
- Implementation of policies should be monitored and researchers' compliance verified through annual performance reports of both researchers and archives.

[7](Accessed 10 April 2018).

Data management does not happen automatically; researchers must be trained in data management techniques. This can be addressed by introducing training through university-level courses in data management and archiving. Field methods courses might include an introduction to workflow management, metadata, access levels, ethical considerations, licensing, and informed consent. Archives could also develop online resources, including video tutorials, in order to ensure thorough coverage of the ethical and practical issues involved. Training for archivists should cover legal and ethical issues. Textbooks and other materials should be developed to allow this, with funding allocated for their creation (see §5).

With respect to the fundamental issue of funding for archiving and making research data accessible, collaboration between archives on a technical level and the sharing of solutions between institutions can minimize costs. Archives need to assess the true costs of curation and archiving, taking into account ingestion, curation, loading, storage, managing access regulations, agreeing on access with speaker communities, and so on, and must seek appropriate sources of funding. Like individual researchers, institutions must be aware of legal requirements regarding making materials available. Researchers need to understand the costs of curation and archiving, and must work with funders to find ways of continuing to fund these beyond the timespan of a grant.

5. Archives Archives as institutions are discussed above in §4. This section examines archives with respect to access, focusing on data types and users. Archives play a critical role in public access to research material, as it is through archives that materials are made discoverable and accessible. While depositors may be better prepared to curate their materials, in practice this task ultimately falls to the archive, which has responsibility for the curation and long-term storage of materials.

5.1 Data types, access conditions, and public access As discussed in §2, providing access to certain types of data may be problematic. A variety of data types are listed in Table 1, together with issues that they may face and possible strategies for dealing with the challenges.

As indicated in Table 1, most material can be made Open Access or accessible through log-in, while access conditions may be appropriate for sensitive material, according to the direction of the speaker or their community. In some cases anonymization may provide a solution, with the researcher undertaking the anonymization with the assistance of archival staff. Metadata can indicate that participants should not be identified: they can be referenced as "anonymous", or people and locations can be given pseudonyms.

5.2 Archives and their depositors This section addresses technical aspects of depositing in archives; see §6 on more personal aspects.

Table 1. Issues and solutions for different data types

Data type	Issues	Strategies
Descriptive metadata	Unproblematic in most cases.	Participants in recording sessions and their personal details as well as locations can be anonymized if necessary. Metadata sets can be hidden while collections are in construction.
Child language data	Minors are protected by national and international laws. It may be necessary to restrict access to voices and images. Consent given by legal guardians may require renegotiation once children come of age; provision must be made for obtaining children's consent later on.	Metadata and anonymized transcripts may be made available. Materials can be archived with restricted access for research use. Video and audio can be stored offline (mandatory in some countries for data pertaining to children).
Original texts, transcripts, and annotations	Less personally identifying than audio/video/images. Intellectual property rights must be respected. Some content may be problematic (see §2.2 on avoidance of harm).	Texts, transcriptions, translations, and some tabular data may be made available where other media are restricted. Can be anonymized. Certain content may need to be restricted. Redacted texts could be made publicly accessible. A limited embargo period may be permitted for students or for first use by researchers.
Multimedia (audio, video)	Contains personally identifying information. Various potential consequences for speakers and communities.	May need to be restricted. Can be made available, if personal rights are cleared and intellectual property rights respected.
Experimental data	Generally unproblematic. Already anonymized.	Existing guidelines from APA, university ethics committees, etc. must be respected.
Location data	Geographical coordinates of certain objects, events or natural resources may be commercially interesting (loggers, poachers, mineral prospectors, bio-pirates, etc.) and may put the community and their area at risk.	Restrict any information that is likely to be problematic. Provide mediated access, if there is a possibility of inappropriate use of the information. Consider withholding from archival collection, if accidental release of data would prove irreversibly problematic.
Sensitive material	Potential monetary value (e.g., ethnobotanical material)	Can be made available to registered users, with clear guidelines for usage and a clear trail of use.
Legacy materials	Not easy to determine access restrictions as there is often no indication of informed consent or sensitivity.	The default is for such data to be publicly available, unless there are legal restrictions or concerns around sensitivity. It should be acknowledged that the material has unclear copyright conditions and can be taken down, if anyone is aggrieved by it (the 'takedown principle', cf. e.g., Urban, Karaganis, & Schofield 2017). Crowdsourcing may be used to enrich metadata and identify possible access issues.

5.2.1 The challenges Archives rely on depositors as intermediaries between themselves and communities, for obtaining informed consent and providing metadata, licenses, and access restrictions. This reliance on the depositor can create problems regarding the handling of personal rights, traditional knowledge, and copyright and licensing rights, especially with older collections where a depositor is no longer available or has not nominated a legal successor to make decisions for the collection, does not have a long-term relationship with the speakers, or where informed consent has not been obtained.

5.2.2 Possible strategies Clear statements of rights and licenses and unambiguous access conditions are crucial for archives to be able to implement the intentions of individuals and communities. From the outset of a project, researchers should work with archives to address issues of licensing and access, to develop a succession plan stating who will be responsible for materials in the future, and to make plans for future treatment of restricted materials. While restricted materials are generally not favored by archivists, community wishes regarding access restrictions must be respected. At the same time, it is too easy for researchers or archives to use 'community sensitivity' as an excuse for not making their records available, resulting in the age-old colonial extraction of materials that do not then find their way back to the source community. In a reflective review of the relationship between Indigenous Knowledge and Open Access, Christen (2012:2889) concludes:

> Incorporating a wider range of ethical and cultural concerns into our digital tools subverts the narrow notions of information freedom and the cultural commons that presently characterize our discussion of the commons. Memes like 'information wants to be free' and general calls for 'open access' undo the social bearings of information circulation and deny human agency. Shifting the focus away from information as bits and bytes or commodified content, indigenous cultural protocols and structures for information circulation remind us that information neither wants to be free nor wants to be open; human beings must decide how we want to imagine the world of knowledge-sharing and information management in ways that are at once ethical and cognizant of the deep histories of engagement and exclusion that animate this terrain.

Archivists can provide guidance and training on obtaining informed consent for archiving as part of creating a data management plan. Clear rules should set out what is expected in terms of access regulations; these might include embargo periods for which any access restrictions must be properly justified. Otherwise, materials for which no justification for an embargo is provided should be made accessible. Funders and archives might share information about depositors' track records on access and archiving.

5.3 Archives and their users

5.3.1 The challenge Language archives must be designed to meet the needs of a variety of users with different expectations and requirements, and these expectations and requirements may change with time (cf. Wasson et al. 2016). Users may include the following:

- Scientific researchers, both in linguistics and other fields, e.g., ethnography, history, cognitive science. They require: good access to data, including detailed search options; streaming and download options; easy ways to reference specific data; ability to upload new annotations without compromising existing ones.
- Speakers of the language and community members. They require: an interface in an appropriate local and/or national languages; metadata and transcriptions in a national language; search capabilities for individuals, places, types of recordings, etc.; an interface suitable for use in schools and other community contexts. Parts of the collection may be accessible only to the community or only to individuals in the community.
- General public, museums etc. Materials and resources that are particularly accessible and interesting, often for extraneous reasons, can be highlighted as "showpiece of the month", etc.; interfaces and transcriptions can be in global languages other than English; holdings described in the language of the general public; links to and from Wikipedia articles and other collaborative platforms.

5.3.2 Strategies to address needs of different user types Different users may have different access rights. For instance, access might be by log-in via a client certificate-based authentication and/or Shibboleth for scientific researchers, and there might be parts of the collection that are restricted in use and available only to community members, or perhaps only to selected community members. Other parts of the collection might be open to all.

Public access includes access to materials by the speakers and their community. Community access deserves somewhat more attention than it currently receives, and can be affected by a variety of factors in different regions. Archived records may not be findable by speakers for a variety of reasons, including:

(a) language barriers;

(b) lack of bandwidth/internet access;

(c) speakers/community not being aware that recordings exist or are available;

(d) inaccurate metadata;

(e) lack of technical skills and computer literacy; and

(f) an interface and data structure that is difficult to use.

Such issues can be addressed by publicizing archive metadata through local cultural agencies and other institutions (e.g., schools, museums, local government), and working to improve access to archive sites. The interface, minimally the metadata catalogue, can be provided in a local language and appropriate training offered. If people do not have access to the internet or computers, tablets or notebooks can be set up in a school or other institution as a local archive. Funders could cover reasonable costs for capacity building and providing local access, with these being implemented by the researcher, the archive, or both, depending on the situation. This must include ongoing training, and to be effective, researchers should work with communities to understand and implement their perspectives on what is needed. Periodic reviews of ownership and access conditions by all relevant parties will likely be helpful. It is important to keep in mind that there is no one-size-fits-all solution – there is both regional variation and variation over time (including changes in technology and in community access to and ability to use technology).

The work of documentation has the potential to be expropriative – collecting and disseminating recordings of indigenous people speaking in their languages is problematic. As Smith (1999:99) notes, "[i]ndigenous knowledges, cultures and languages, and the remnants of indigenous territories, remain as sites of struggle".

However, archive work is typically driven by non-indigenous university-based researchers who have taken on the responsibility of making the research of the university available outside academia. This action counteracts an earlier expropriation, that of the academic researcher who kept recordings safe but did not know how to return them to the source communities, or did return them periodically on analog media that had a short life span.

5.4 Embedding in institutions Some archives are embedded in larger institutions (as opposed to community-based archives, for example) and must follow internal policies, including internet security protocols, choice of specific models and systems for archiving. While institutional policies may conflict with various archival practices, we suggest a commitment to provide public access should form a general archiving principle. Note that prior agreements with depositors may be legally binding; for instance, access levels and other similar requirements need to be preserved.

6. Credit, control, and public access Concerns within communities about making data public were addressed in §2 and §5.1. This section addresses concerns by researchers about making data public.

6.1 Credit, control, and the researcher This section addresses two concerns of researchers: (1) identifying who should be attributed credit may be difficult, or contested; and (2) researchers or research teams may be ambivalent about making a collection available as they are concerned that their contribution to gathering, transcribing, glossing, and translating the material will go unrecognized. We focus on credit

with regard to researchers and communities. Funders are generally acknowledged in a footnote rather than through authorship (we recommend footnote acknowledgement of funding for all archived collections as well as for publications).

Documentation teams should discuss who will be credited in references to the data collection, and how. Major language consultants (transcribers, translators) might be included in references to the whole collection, while individual speakers who contribute narratives, songs, etc. might be credited only in the metadata for individual sessions. The entire team needs to understand the different contributions and what they involve in order to make such decisions – this might come about through workshops revolving around issues of consent. We recommend that the relative contributions of individual contributors are explicitly described in data collections.

In publications arising from language collections, each individual's contribution must be considered when determining co-authorship versus acknowledgement. The relative contributions of individual contributors should be explicitly described in the publication.

Research teams should do what they can to make credit by citation easy. Creators of collections should provide explicit and easy-to-find citation guidelines with the collection (with archives providing guidelines for citing whole deposits, as well as data and metadata at more granular levels; see for example the citation guidelines provided by AILLA at https://ailla.utexas.org/site/rights/citation). Users should cite examples by giving proper references, and researchers who make substantial use of particular collections for a publication should consider including the compilers as co-authors. Compilers of data collections can present the structure of their archival deposit in a journal publication (e.g., Salffner 2015; Caballero 2017; Oez 2018) as a citable reference to the collection. Archival resources can also be cross-referenced in collections such as Glottolog.[8]

6.2 Credit, control, and access restrictions Access restrictions were mentioned in §5.1, and we return to them now, first looking at access restrictions and the community, and then at access restrictions and the researcher. We continue to draw a line between community and researcher, although in reality such lines can blur.

6.2.1 Credit, control, access restrictions and the community Language documentation typically works with languages spoken by a small number of speakers. Due to the small size of the cohort, recordings can contain materials which might put these communities at risk of harm, from outside or from within. A text might cause harm by asserting the rights of a particular group to a contested piece of land or a favorable version of history. Other recordings contain highly personal information, and in small societies it may be impossible to anonymize speakers.

While funders may require public access, community members may require restrictions before information is provided. Sensitive materials archived with restrictions can at least be preserved. Some researchers find setting immediate restrictions may

[8]http://www.glottolog.org. (Accessed 21 May 2019).

lead to Public Access over time, as people decide that they want materials to be accessible.

Where not at odds with the community's views, we recommend using restricted access only with a clearly specified embargo period, after which the restrictions can be lifted. That date could possibly be in the far future, but it must not be undefined. For any materials requiring long-term restrictions, legal successors to depositors should be identified wherever feasible (this implies an ongoing relationship at least between archives and researchers).

6.2.2 Credit, control, access restrictions, and researchers Researchers may avoid making their data collections publicly available out of fear that others might use the data without proper attribution. Creators of research data have a recognized right to reasonable first use of data. It is therefore possible to restrict access to data collections/corpora for a defined period to enable primary compilers to work with their data before others do (cf. Berez-Kroeker & Henke 2018:362–364). However, embargo periods should not be perpetuated without limits. Archives should require justifications for extensions beyond a standard embargo period (see §4.2). The risk in not allowing material to be embargoed is that not all records will be archived and they will then potentially be lost. Once data is released, citation standards for data sources must be applied and checked/enforced by peers and peer review processes when it is observed that data is being reused (see §6.1).

7. Summary This paper discusses some of the challenges arising from the ideal of Open Access to collections that result from language documentation projects. These include challenges involving communities, legal matters, archiving, costs, data types, access types, and credit. This paper suggests some possible solutions, noting the importance of being aware that communities, data contexts, and technology all evolve over time. In all areas, we emphasize the need for learning what external forces there are that must be complied with, and for focusing on education, on working together, and on flexibility at all levels.

Acknowledgements This paper is the result of a 3-day workshop on "Open Access and Open Data of Endangered Languages Collections" held October 10–12, 2016, at the University of Cologne funded by the Volkswagenstiftung. It emerged as collaborative writing by 35 stakeholders and researchers working on different aspects of language documentation and Open Access. Thanks to two anonymous reviewers for their constructive comments.

References

Anderson, Jane & Molly Torsen. 2012. *Intellectual property and the safeguarding of traditional cultures: Legal issues and practical options for museums, libraries and archives*. Geneva, Switzerland: WIPO. http://www.wipo.int/edocs/pubdocs/en/tk/1023/wipo_pub_1023.pdf.

ATHENA. 2009. *ATHENA deliverables and documents: WP6 – Analysis of IPR (Intellectual Property Rights) issues and definition of possible solutions*. http://www.athenaeurope.eu/index.php?en/149/athena-deliverables-and-documents.

Austin, Peter K. 2010. Communities, ethics and rights in language documentation. In Peter K. Austin (ed.), *Language documentation and description*, vol. 7, 34–54. London: The Hans Rausing Endangered Languages Project.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard Meier, Nicholas Thieberger, Keren Rice, & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 57(1). 1–18. doi:10.1515/ling-2017-0032.

Berez-Kroeker, Andrea L. & Ryan Henke. 2018. Language archiving. In Rehg, Kenneth & Lyle Campbell (eds.), *Oxford handbook of endangered languages*, 347–369. Oxford: Oxford University Press.

Bhattachary, Darren & Douglas Dalziel. 2012. Open data dialogue: Final report. *Research Councils UK*. https://www.ukri.org/files/legacy/documents/tnsbmrbrcuk-opendatareport-pdf/.

Bowern, Claire. 2010. Fieldwork and the IRB: A snapshot. *Language* 86(4). 897–905.

Brown, Michael F. 2003. *Who owns native culture?* Cambridge, MA: Harvard University Press.

Caballero, Gabriela. 2017. Choguita Rarámuri (Tarahumara) language description and documentation: A guide to the deposited collection and associated materials. *Language Documentation & Conservation* 11. 224–255. http://hdl.handle.net/10125/24734.

Choukri, Khalid, Stelios Piperidis, Prodromos Tsiavos, Tasos Patrikakos, Maria Gavrilidou, & John Hendrik Weitzmann. 2012. *META-SHARE: Licenses, legal, IPR and licensing issues*. Berlin, Germany: META-NET. http://www.elra.info/media/filer_public/2015/03/30/meta-net-d613.pdf.

Christen, Kimberly. 2012. Does information really want to be free? Indigenous knowledge systems and the questions of openness. *International Journal of Communication* 6. 2870–2893. https://ijoc.org/index.php/ijoc/article/view/1618.

Christen, Kimberly. 2015. Tribal archives, traditional knowledge, and local contexts: Why the "s" matters. *Journal of Western Archives* 6(1). Article 3. http://digitalcommons.usu.edu/westernarchives/vol6/iss1/3.

CLARIN (Common Language Resources and Technology Infrastructure). Licenses and CLARIN categories. https://www.clarin.eu/content/license-categories. (Accessed 10 April 2018).

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Gippert, Jost, Nikolaus. P. Himmelmann, & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Walter de Gruyter.

Fluehr-Lobban, Carolyn. 1994. Informed consent in anthropological research: We are not exempt. *Human Organization* 53(1). 1–10. doi:10.17730/humo.53.1.178j-ngk9n57vq685.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, & Tyler Heston. 2017. Putting practice into words: Fieldwork methodology in grammatical descriptions. *Language Documentation & Conservation* 11. 157–89. http://hdl.handle.net/10125/24731.

Good, Jeff. 2018. Ethics in language documentation and revitalisation. In Rehg, Kenneth & Lyle Campbell (eds.), *Oxford handbook of endangered languages*, 419–440. Oxford: Oxford University Press.

Grinevald, Colette. 2006. Worrying about ethics and wondering about "informed consent": Fieldwork from an Americanist perspective. In Saxena, Anju & Lars Borin (eds.), *Lesser known languages in South Asia: Status and policies, case studies and applications of information technology* [TiLSM 175], 339–370. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196.

Janke, Terri. 1998. *Our culture, our future: Executive summary of report on Australian Indigenous Cultural and Intellectural Heritage Rights*. Canberra: AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies) and ATSIC (The Aboriginal and Torres Strait Islander Commission). https://www.wipo.int/export/sites/www/tk/en/databases/creative_heritage/docs/terry_janke_culture_future.pdf.

Khan, Mehtab. 2018. *Traditional knowledge and the commons: The open movement, listening, and learning*. Creative Commons [blog]. https://creativecommons.org/2018/09/18/traditional-knowledge-and-the-commons-the-open-movement-listening-and-learning/.

Klimpel, Paul. 2013. *Free knowledge based on Creative Commons Licenses: Consequences, risks and side-effects of the license module "non-commercial use only – NC"*. Berlin: Wikimedia Germany. https://openglam.org/files/2013/01/iRights_CC-NC_Guide_English.pdf.

Macmillan, Fiona. 2013. The protection of cultural heritage: Common heritage of humankind, national cultural "patrimony" or private property? *Northern Ireland Legal Quarterly* 64(3). 351–364. http://eprints.bbk.ac.uk/7289/1/7289.pdf.

MINERVA EC Working Group "Quality, Accessibility and Usability" (ed.). 2008. *Handbook on cultural web user interaction*. 1st edn. http://www.minervaeurope.org/publications/Handbookwebuserinteraction.pdf.

Næss, Åshild & Even Hovdhaugen. 2011. Language is power: The impact of fieldwork in community politics. In Haig, Geoffrey, Nicole Nau, Stefan Schnell, & Claudia Wegner (eds.), *Documenting endangered languages: Achievements and perspectives*, 291–304. Berlin: Mouton de Gruyter.

Newman, Paul. 2011. Copyright and other legal concerns. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 430–456. Oxford, New York: Oxford University Press.

Nowviskie, Bethany. 2014. *Why, oh why, CC-BY?* Bethany Nowviskie [blog]. http://nowviskie.org/2011/why-oh-why-cc-by/.

OECD (Organisation for Economic Cooperation and Development). 2015. Making open science a reality. *Science, Technology & Industry Policy Papers* No. 25. Paris: OECD Publishing. doi:10.1787/5jrs2f963zs1-en.

Oez, Mikael. 2018. A guide to the documentation of the Beth Qustan dialect of Central Neo-Aramaic language Turoyo. *Language Documentation & Conservation* 12. 339–358. http://hdl.handle.net/10125/24773.

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

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

Rice, Keren. 2011. Ethical issues in linguistic fieldwork. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 407–429. Oxford, New York: Oxford University Press.

Robinson, Laura C. 2010. Informed consent among analog people in a digital world. *Language & Communication* 30. 186–191. doi:10.1016/j.langcom.2009.11.002.

Rundle, Hugh. 2014. *Creative commons, Open Access, and hypocrisy*. Information Flaneur: Hugh Rundle [blog]. https://www.hughrundle.net/2014/03/24/creative-commons-open-access-and-hypocrisy/.

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

Schmidutz, Daniel, Lorna Ryan, Anje Müller Gjesdal, & Koenraad De Smedt. 2013. *Report about new IPR challenges: Identifying ethics and legal challenges of SSH Research*. Deliverable D6.2 of Data Service Infrastructure for the Social Sciences and Humanities (DASISH). http://dasish.eu/publications/projectreports/D6.1_final.pdf.

Selfe, Cynthia. L. & Gail E. Hawisher. 2004. *Literate lives in the Information Age: Narratives of literacy from the United States*. Mahwah, NJ: Lawrence Erlbaum Associates. doi:10.4324/9781410610768.

Singer, Ruth. 2014. *Open access and intimate fieldwork*. Endangered Languages and Cultures [blog]. http://www.paradisec.org.au/blog/2014/03/7940/.

Smith, Linda Tuhiwai. 1999. *Decolonizing methodologies: Research and Indigenous peoples*. London, New York: Zed Books; Dunedin, New Zealand: University of Otago Press.

Suber, Peter. Last revised 2009. Timeline of the Open Access Movement. https://legacy.earlham.edu/ peters/fos/timeline.htm. (Accessed 21 May 2019).

Thieberger, Nicholas. 2014. The cost of not archiving. Presented at the 3rd InNet conference, Budapest, Hungary, September 5–6. http://www.nthieberger.net/CostOfNotArchiving.pdf.

Thieberger, Nicholas & Simon Musgrave. 2007. Documentary linguistics and ethical issues. In Austin, Peter K. (ed.), *Language documentation and description*, vol. 4, 26–37. London: SOAS. http://www.elpublishing.org/PID/048.

Tyner, Kathleen R. 1998. *Literacy in a digital world: Teaching and learning in the age of information*. Mahwah, NJ: Lawrence Erlbaum Associates.

Urban, Jennifer M., Joe Karaganis, & Brianna Schofield. 2017. Notice and takedown in everyday practice. *UC Berkeley Public Law Research Paper No. 2755628*. doi:10.2139/ssrn.2755628.

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.

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

Whimp, Kathy & Mark Busse (eds.). 2000. *Protection of intellectual, biological and cultural property in Papua New Guinea*. Canberra: Asia Pacific Press.

Wilkins, David. 1992. Linguistic research under Aboriginal control. *Australian Journal of Linguistics* 12. 171–200.

WIPO (World International Property Organization). 2016a. *Traditional knowledge and intellectual property*. Background Brief No. 1. http://www.wipo.int/edocs/pubdocs/en/wipo_pub_tk_1.pdf.

WIPO (World International Property Organization). 2016b. *Documentation of traditional knowledge and traditional cultural expressions*. Background Brief No. 9. http://www.wipo.int/edocs/pubdocs/en/wipo_pub_tk_9.pdf.

Mandana Seyfeddinipur
ms123@soas.ac.uk