The following full text is a publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/122542

Please be advised that this information was generated on 2020-03-08 and may be subject to change.
An analysis of research implications regarding digital archives of spoken language

In the framework of programmes for documenting endangered languages such as those funded by the Volkswagen Foundation and the Arcadia Fund, unprecedented amounts of audiovisual data on endangered languages and cultures from around the world are currently being electronically archived. The expectation is that the materials collected will be more readily available (and for much longer) than previously, and available in ways that would benefit a number of different groups of potential users, including speakers who want to revitalise their languages and cultures. However, as I have argued elsewhere (Widlok 2010), the new electronic archives are not simply a quantitative extension of existing modes of data collection, but they qualitatively alter the relationship between researchers and their products and, as a consequence, also the relations between the researched and these products and the relationship between researchers and their partners in the field. The new possibilities of Internet-based digital archiving and online databases are much more than "just technical" innovations. Rather, the new archiving technology is also changing the ways in which we generate and share knowledge. The first half of this paper, therefore, aims to lay out
in some detail what is implied in the broad processes of electronic data gathering, digitisation, and online archiving.

Breathing life into data cemeteries

One of the most prominent assurances of new digital archiving technologies is that it can help to prevent data loss and data cemeteries. In the past, many recordings of spoken language have effectively been lost, not only materially but also through being buried in personal archives. The problem continues into the present and, arguably, it has been aggravated since the costs of recording have dropped dramatically. For myself and many colleagues, a major incentive to engage with digital archiving was to seek a strategy for coping with an ever-increasing private collection of audio- and videotapes, originating from various research projects over the years, materials for which it became ever more difficult to find a machine that would allow the data to be used in the future. Increasingly, there are also recordings, usually audio tapes, produced and kept by members of the speech community, but they frequently get recorded over after a while or are lost in closed collections. Field researchers tend to keep their data, but this data often ends up in data cemeteries, boxes of tapes awaiting transcription and translation. The prospects for these private collections are rather bleak: data loss due to media deterioration or due to the decommissioning of research projects and careers. My personal motivation for getting involved with the endangered languages programme DOBES (Dokumentation bedrohter Sprachen) of the Volkswagen Foundation, was the hope that the audio-visual data that I had accumulated in years of field research with ≠Akhoe Hai//om in northern Namibia could be prevented from deterioration and could be made accessible to others. After several years of running the project, many of my old tapes are now digitised and archived, but at the same time many more tapes have been added so that the total amount of untranscribed and unanalysed data is actually greater than what I started with. Moreover, the data collection has changed in ways that I had not anticipated when I began digital archiving. While I initially treated archives as passive “dumping grounds” it soon became clear that the digital archive was striking back, prompting me to reconsider basic assumptions and to change some of the habitual ways of doing research. The storage technology changed the record and changed the method of data gathering in at least three different ways, namely in terms of a departure
from earlier holistic approaches, by fostering modularity and through the introduction of standardised metadata.

**Farewell to holism**

The composition of teams that make up documentation projects funded by the research initiatives already mentioned above is typically heterogeneous. The interdisciplinary teams include anthropological and linguistic researchers at a number of different levels (post-doc, PhD, research assistants, interpreters etc.) and for different time periods (short contracts, PhD projects, and as part of lifetime engagements). Since all members contribute to the data collection in different ways and at different times, the result is an open corpus very unlike the typical ethnographic monograph or conventional linguistic collection of folk-stories that tend to be holistic if not in scope then at least with regard to the fact that they are presented as a book or a similarly bounded entity. Ideally, the new digital corpora are supposed to grow even after the funding period has ended as researchers add to it and work on it for their various projects, and to various degrees this ideal is in fact realised. The corpus is eventually shaped not only by the original team of researchers but also by collaborators and by interested colleagues whose actions shape at least part of the corpus as they use it for a variety of purposes. This new set-up destroys any illusion that one might have had as the author of an ethnographic monograph or linguistic text collection in terms of holistic completeness and closure. On the positive side of things, the digital corpus more honestly reflects the fact that any field research is a long-term process of accumulating and revising knowledge, a process that tends to be hidden in the production of books and volumes. However, it remains to be seen how many of the newly established corpora will indeed become living bodies instead of turning into yet larger data cemeteries. Given that the funding agencies guarantee to keep the data collection active for decades (by migrating data to readable formats in the future), the chances are that we will see some interesting developments concerning the social life of data files. An inevitable trade-off in this development is that researchers partly lose control over the end product and, potentially, so do their informants. The researcher may be able to discuss publications with informants before they get printed, and will seek consent before things are recorded and put into an archive, but there is no way that a constantly growing and
changing body of data that is subject to collective work and revision could be controlled in the same way. I shall return to this point in the second half of this chapter.

Welcome to modularity

It is tempting to look at digital archives as open-ended corpora that do away with the limitations of former collections of texts as indicated above. However, this primary openness should not be confused with amorphousness. For an electronic archive of spoken language to work, at least in the current set-up provided by electronic archiving software, it has to be organised in modules, usually called “sessions”. These sessions are the basic units of data storage in electronic corpora. They can range from tiny one-sentence recordings to hours of videoed ceremonies and story-telling events. Moreover, the system allows for (and even encourages) any one recording to be organised in more than a single session, but the recording must minimally be part of one session (with metadata) in order to be visible in the corpus. The hypertext structure of the archive allows the underlying digitised tapes to be cut and joined in as many different ways as those working with the corpus care to specify. In our own project, for instance, we envisage that a healing dance that lasts several hours will form one session as a dance event, but that sections from this dance may also form separate sessions. In another example (see Widlok 2010: 51), a folk storytelling event may form a session of trickster stories, of cooperative story telling or of mother-in-law taboo since the underlying event is each of these three, namely a young man and his mother-in-law jointly narrating a trickster story.

Moreover, some projects have already experimented with community platforms that sort and present the data sessions not in terms of the categorical system of comparative linguistics or comparative anthropology, but in terms of what community members have found to be a useful organisation for their specific local purposes. In our own project, we have seen the beginnings of this in the use of sessions for community purposes: a common problem with largely egalitarian groups such as the ≠Akhoe Hai//om is that members do not easily speak for the community. When they get invited to represent their community at meetings in the capital, they tend to say little out of fear of getting criticised for what they have said when they return. This uneasiness to speak for others can be softened by using electronically archived materials. At one recent NGO-run workshop in the
capital, the group members attending were showing video-clips (sessions from interviews) with voices from their home place. This allowed for more voices to be heard at the meeting without the delegate being forced to represent others in the context of their own established social practice that does not operationalise representing others as it is required by government or non-governmental organisations. Until recently, anthropologists (or other intermediaries) have often felt pressed to take over the role of speaker for the local community in these situations. Now they can take on the less contentious role of facilitators who only provide the technical equipment for what one may call a local appropriation of the archived materials.

The versatility of sessions as data clips is well established since YouTube and other Internet platforms have become ubiquitous sources. What is occasionally overlooked is the extent to which the often diffuse complexity of speech events and of ethnographic situations in which language is spoken becomes modularised with hard and fast boundaries so that we take these models to be true representations of the events from which they originate. In other words, there is a danger that the holistic illusion of a complete corpus that is dismantled in electronic corpora (see above) gets re-established at the level of sessions. In principle, there is no reason why we should not continue to cut new sessions from the original digital media files as we continue working, but in practice, the sessions—once established—tend to take on a life of their own.

**Better data with meta-data?**

Just as an organisation of the corpus in terms of sessions seems inevitable in the context of digital archiving, the same holds for the creation of metadata files that organise the data itself. To begin with, the sheer size of the data collections discussed here makes metadata essential. There is no easy way to find relevant cues if one has to go through hours of audio or video recording unless there is metadata that provides hints on the contents of the recording. Other minimally required metadata includes information about who collected, cut and processed the recording. Metadata is the main channel whereby context is preserved in terms of who said what, to whom, and in what kind of setting. As soon as the material leaves the confines of a private archive under the control of a researcher who can comment on the circumstances and details of the recording, and as soon as it reaches a public archive with considerable longevity, metadata is critical for contextualisation.
Correspondingly, archivists present metadata as the most critical resource for preventing the small data cemeteries mentioned above from being merged into huge data cemeteries. At this stage it becomes clear that, even with direct community involvement in compiling the corpus, digital archiving does not comply to the ideal of a dialogical research exercise with no power differentials. Clearly, the researchers who formulate the metadata (specifying participants, location, genres etc.) and the archivists who provide the templates for the metadata (controlled vocabularies, drop-down menus, boxes to be filled in) very much determine how sessions are described and corpora are compiled. Having said that, this of course is also true for many, if not all, data compilations that field researchers have hitherto come up with, be it text, audio, video or other. Context can never been exhausted, and there is always a selection of context. Whatever effort is made to make local voices heard, choices about the compilation and composition of the record usually remain with the researcher. The difference, compared to earlier practice, is that the metadata requirements of electronic archives render it necessary for the researcher to make his or her system of contextualisation open and explicit; and to agree with others on some shared standard. We all categorise our information (and to some extent our informants do that, as well) in one way or another, and similarly the events and situations from which the data is being derived. The metadata files in the corpora discussed here make it necessary to be explicit about these categorisations. Many researchers who are devoted to a particular language (and language community) are uneasy about the standardised categorisations of metadata descriptions. In any case, metadata specifications provide the opportunity to reflect on these categories. A major gap in the metadata that we found in our own project (see Widlok, Rapold and Hoymann 2008) is that the person-related information is usually individualised. Considerable effort is made in electronic archives to allow for a number of ways of anonymising speakers as individuals. However, the effort to connect the person-related information into a network (for instance of kin-relations between speakers) is still in its infancy. The metadata currently provides a purely summative list of informants and, as yet, no knowledge about the participatory frameworks in which data was generated.\footnote{For the importance of participatory frameworks for understanding spoken language see William Hanks (1996: 142).} In other words, research projects at present give information about individual contributors, but not about the social links between
them. “KinOath”, a new piece of software currently in development by Peter Withers at the Max Planck Institute for Psycholinguistics in Nijmegen, aims to close that gap. It is closely integrated with the “Arbil” metadata management tool.

**An analysis of property and access rights regarding digital archives of spoken language**

Digitisation not only affects the process of data collection and research but also the possibilities and limitations of access to that data. I am here not concerned with the (important) questions of access to digital technology and Internet connection—which continues to be a problem in many parts of the world—but rather those issues of regulating and managing access that, typically, linguists and anthropologists are expected to solve. Technically, the question of access rights to the collected data is solved in the sense that the archive allows researchers to categorise their data into different levels of access. Funding agencies like the Volkswagen Foundation would like the default level to be set at open and public so that all data (and not just the metadata) is openly accessible unless a specific reason is given to make it a temporarily restricted source. In fact, most project teams tend to see it the other way around. They open up very few selected show pieces for which there is open access. The largest part of the body is limited access with two thresholds: the lower threshold consists of an automatised declaration with which the user must agree along similar lines as agreements to Internet downloads and other web services. Instead of agreeing to a licence agreement, the user here agrees to have read the DOBES code of conduct and to comply with it in terms of protecting local communities and their intellectual property rights. The next threshold is that interested users have to get in touch with a responsible corpus manager (typically a member of the research team) who can then advise the archivist to grant them access to selected sessions from the corpus. Typically, this allows fellow researchers to make use of the corpus while the research team maintains some control over who uses the data. Closed access only exists as a temporary measure that is taken to grant researchers, especially PhD students, a period of exclusive use rights until they have completed their degree or publication. As I have pointed out elsewhere (Widlok 12–13 June 2008), these uniform and clear-cut technical access categories are in stark contrast to issues of
property rights that are in most cases overlapping and of access rights that are typically shared, especially with regard to the long-term perspective.

The point to highlight here is that what changes with regard to earlier processes of granting or delimiting access is that we are no longer dealing with a dyadic negotiation between researcher and researched. Rather, there are now more parties involved: the funding agencies, with their open access policies but also, potentially, community agencies, often following restrictive practices and potentially in conflict with one another about who is the legitimate holder of the rights of the community. This raises the thorny issue of “group rights” and “cultural property” (see Barry 2001). We have seen conflicts of this sort arise over the repatriation of ancestral bones and artefacts to present-day indigenous communities and this suggests that similar problems may arise with electronically archived materials, especially when individual informants (and the responsible researchers) are no longer alive. While there is a need to specify access rights in the metadata, it is an illusion to think that this sufficiently accommodates for the complexity of rights issues. A first step towards solving many questions about property, I maintain, is to recognise at what level rights are actually claimed. In most cases property is made up of a bundle of rights (of ownership, of access, of use, of alienation, of inheritance) so that the recognition of authorship need not imply rights at other levels (as any author of a scientific publication knows).

A layered model of property and access rights

Many conflicts and misunderstandings surrounding property rights in data result from the fact that different layers of property rights are not sufficiently distinguished. For the purposes of this chapter we may distinguish layers concerning 1) values, 2) regulations, 3) relations, and 4) practices. For instance, we may all agree on certain values (such as the protection of privacy and the openness of scientific results) but there may still be debate about what kinds of regulations are best suited to realise these values, especially if more than one value is to be considered. Apart from that, there are discrepancies that may arise between different layers that make up complex property rights. I therefore suggest to apply to digital archiving of spoken language a layered model of property relations that has been developed in legal anthropology (see Benda-Beckmann and Benda-Beckmann 1999; and Widlok 2001). This model looks at property as
a bundle of rights at the level of values, regulations, relations and practices. While language documentation programmes tend to be fairly outspoken about layers of values and regulations, they are less explicit about layers of social relationships and practice.

**Cultural values of access**

In a sense, cultural values of access is the most unproblematic layer of property rights, since there are a number of existing documents—developed after long discussions within the scientific community—that can be referred to. The DOBES Code of Conduct (CoC) has been discussed extensively in this regard; it also includes references to other existing codes. Note that there are tensions built into all these codes since some values are at least potentially in conflict with one another, e.g. the right of privacy and the right of access to scientific results. This is the normal state of affairs and therefore we should not shrink back from embracing these values, even if they are partly in conflict with one another. This is why we need to look at the other layers, which can take the sting out of these tensions.

Each individual documentation team can refer to the relevant values that are formulated in documents such as the code of conduct, i.e. the respect towards intellectual and cultural property rights, the privacy of individuals, and the obligation to make the material accessible to interested non-commercial uses. However, even if we assume that there will be no changes in these values in the long-term future, inevitably, existing tensions between values will be resolved in different ways at various points in time. The current tendency is clearly towards open access. However, with increasing commoditisation, and possibly with the increasing disappearance of languages, the tendency may shift towards restricting access. The community of researchers now considers cultural heritage a treasure, but to some it may become a burden, too. The value of archiving may itself change in the future. It is not possible to foresee these developments, since our knowledge increases as language documentation grows. With spoken language materialising into data sessions and corpora

---

2 For a list of relevant documents, see International Association of Sound and Audiovisual Archives (IASA)’s *Copyright & Other Intellectual Property Rights* and The Hans Rausing Project for Endangered Languages’ *Online Resources for Endangered Languages. Ethical issues* in Online Sources.
we may securely assume that the same dynamics are likely to emerge, that anthropology has found to be implicated in “the social life of things” (Appadurai 1986) in other domains of materialised culture. Given the longevity of data collection, many more conflicts become possible that did not apply when records of spoken language were more fleeting. As access rights will no longer be established once and for all but may change, these changes will have to be traced. In concrete terms, this means that a time-tag is added to every step in the digital archiving process, not just when setting up the metadata but with every change made to the access regulations. In an archive of “data objects” that can become subject to conflict, we need to know who made which access decision, when, and for how long it is to be effective.

**Cultural regulations**

The cultural values of access are operationalised into regulations by the teams running digital archives (in consultation with researchers). In the DOBES framework, there are a number of forms and rules such as the Usage Request Form (DOBES-UR), the Depositor-Archivist Agreement (DOBES-DAA) and Usage Declaration (DOBES-UD), as well as the Data Access and Protection Rules (DOBES-DAPR), which cover many aspects of what needs to be regulated with the help of forms.

The forms mentioned comprise some fundamental rules such as “all metadata [and all software] is openly accessible” and “by default, all archived materials [...] are not openly available and access therefore will be restricted” (DOBES-DAPR) as well as the more detailed rules as to what the conditions and procedures are for gaining access. It is noteworthy that at this layer, too, there are inbuilt tensions. For example, with regard to the notion of *shared* property and access rights according to which the copyright rests with both depositors and consultants (e.g. speakers), one can easily imagine situations where among depositors, among consultants or between these groups there may be conflicting interests. The regulations cannot and will not solve this problem for all possible constellations because these very much depend on what happens at the other layers (relations and practices, see below). At this point, the depositors are said to “always have unrestricted access” while the recorded persons have “a right to access” (DOBES-DAPR). In other words, the depositors have the privileged right of access for a maximum of three years and the privilege to formulate
the access rights. The consultants have the privilege to allow or veto any commercial uses of the corpus material (a rule contained in the CoC). These rights are usually already contained in the research agreement that each DOBES project presented before research began. In the ≠Akhoe Hai//om case, this is a contract with one of the local non-governmental organisations. The contract, originally a contract for the media, was adapted in a way that it states:

1. Joint ownership held by the community of ≠Akhoe Hai//om speakers, the individual consultants, and the authors (in resulting publications)
2. The non-commercial nature of the project
3. Reference to the DOBES Code of Conduct
4. The fact that only openly accessible data will be collected

Note that while these regulations clarify the different types of rights that make up the whole bundle of rights that we cover under the term “property”, they often pre-suppose that the holders of these rights are clearly defined. This is the weak point in most regulations of this sort, since the “community of speakers” is a vague concept, at best, and an outright fictional “body”, at worst. Community organisations are known to be highly flexible and prone to conflict, fission and faction fighting. Projects should be aware that “the creators” or “the consultants” are not the same as the organisations with whom the contract has been made. As the case may be, a project team may draw up contracts with different bodies (just as consultants may enter into a contract with more than one research team). In the ≠Akhoe Hai//om case, we have a contract with a national NGO but we could have entered at the same time into contract with the national archive or the national university, and we have tried (so far unsuccessfully) to have yet another contract by creating a local voluntary interest group of people who have an interest in the corpus. If there are conflicts between individuals and organisations (or among organisations), there is no way that any such a contract and its regulations can prevent language documentation to be drawn into these conflicts. The best that one can hope for is that, on the basis of the Code of Conduct, language documentation does not create new conflicts, and that it does not lend itself easily for one faction or interest group to dominate others. The contracts or agreements should not create the illusion that with these regulations in place conflicts are eliminated. Rather, good contracts do not
confuse the rights of all individuals concerned with the specific rights of the representative groups mentioned in contracts. Although a contract is typically a bilateral agreement (in terms of signatures), it can nevertheless be written as a three-party agreement between researcher, counterpart and the speakers, thereby recognising the difference between individual consultants and the organisations that claim to represent the community. In all likelihood, some third party, be it future researchers or speakers, will be affected by the agreements made. Even the so-called “final access statement” (to be made at the end of funding) is not really final insofar as a review of access rights is possible at regular intervals (every two years, see DOBES-DAA). This applies primarily to the distinction into three levels of access (open, on request, not accessible). Passwords that allow access to digital archives are given limited lifetimes (see DOBES-DAPR). Although there is a general tendency in archiving for resources to open up as time goes by (as the collection survives individuals, for instance), there is no categorical reason why access should not also be tightened up after a while, for instance if abuse has been occurring. Researchers and consultants could take advantage of the fact that regulations can be created with some sort of inbuilt shelf life date by which past decisions have to be reconfirmed or revised as a consequence of evolving social relations.

Social relations

It is important to recognise that some aspects relating to property issues cannot be covered with the help of regulations and forms (and in fact need not, or are better not, covered using these forms). For instance, we need to recognise that many statements about property and access rights in digital records of spoken languages are not necessarily about the relation between person or community and corpus. It may have much more to do with the relationship between community and person (e.g. the researcher) or between communities (e.g. speakers of a neighbouring variant). In other words, the lesson from the ethnography of property rights is that people often make property claims not because they necessarily want or need exclusive access to data, a particular recording for instance, but because they want to shape their relationship with others. Property and access claims signal to the rest of the world that local people want to be treated as equal, sovereign and autonomous in their decisions. These social aspirations are only partly satisfied by the use of forms and regulations. More often they are appropriately recognised and satisfied when complemented by other,
more culturally sensitive modes of recognition. In its simplest form these are notes of recognition written by the researchers and covering not just the community of speakers but also, for instance, government ministries that provided research permits and any other stakeholder who has helped in the research. There are other means, some of which are already in use by language documentation projects. Publishing non-academic texts is one such way that shapes social relations. They are not just spin-offs from what we actually do but they are directly implicated in the social relations of property rights, and therefore should be considered an integral part of any language documentation. Another means is the installation of regional data servers which allow communities with restricted Internet access to use the corpus locally and, at the same time, functions as an important step in appropriately locating due recognition.

The recognition of the social relations layer can be both a relief and a challenge. It can be a relief in that it is open to creative ways of recognition beyond the signing of contracts. In some contexts the cultural standing of signed papers may be much lower than (or in any case complemented by), say, the presence of researchers at relevant events and occasions or their engagement for the community of speakers in other ways, for instance in dealings with the media or other outside agencies. At the same time, it can also be a challenge, in that it is simply not good enough to argue “we have signed the appropriate papers” and therefore assume that everything that needs to be done has been done. Contracts defining property rights not only define the relation between people and things (e.g. the corpus), but above all relations between people and other people.

Digital archives with Internet access face the additional practical problem of restricting something once it is out in the public domain. It is quite possible to imagine a situation in which materials, despite being technically accessible, become factually impossible to use because they would harm relationships. For instance, while researchers tend to assume that privacy rights decrease with time (as persons die), this may not be universally true, as—in many other contexts—people do not want unrestricted viewing of images relating to deceased members of their community. Even if copies of the images were already out in the open, it may be considered an impossibility to include them in displays or as examples in publications. More generally, people who may have agreed (even with their signature) to have material on public display may change their mind as they begin to realise what the Internet is and what open access may imply. Although they may not have any contractual (legal) basis for
enforcing restrictions, researchers will no doubt consider accommodating these reservations in order to maintain a good working relationship. At the same time, honouring the decision of a speaker is different from honouring a decision made by a descendent of that speaker. In other words, there is no automatism that would grant a descendent of any contributor to the corpus the same rights as the contributor him- or herself held. They are in different positions to one another and with regard to the corpus. As access rights are handed on (possibly over generations), they may be subject to revision. There have to be default procedures for regulating access in the long term, but these have to be necessarily preliminary in the sense that they do not exist independently of the social relationships that do change over time.

**Social practice**

Finally, there is a layer of property relations which ultimately falls under the responsibility of each researcher and which will vary not just between field sites, but also between researchers. Many researchers who have worked with the same people for two decades have an informal agreement with people in the field. The collaborators in the field, in turn, then have a fairly good idea what the research is all about and are able to clearly signal when to switch off the camera or recorder if they want something to be off record. In fact, many researchers who have reached these means of understanding and agreements, whether tacit or explicit, which argue that—at the layer of practice—these often work just as well or even better than any written form of consent. However, there are other places and other situations, for instance with newcomers to the field or new people to be included as informants, where these agreements do not suffice. The problem is that only those who know the field and the people concerned will be able to tell and to decide on the appropriateness of formal regulations in actual practice.

It is important to note, however, that the practice layer of rights is of equal importance to all other layers. It is obviously not acceptable to use consent forms when one has—in practice—got the signature through subtle forms of cheating or bribery (giving of gifts) or through ignorance (not fully explaining its purpose). But then: who can fully guarantee that the counterpart has fully understood the implications of a piece of spoken language being included in a digital archive that is widely accessible? Are we, as researchers, able to see all future implications? Here it is highly relevant to include in the metadata what the research practice was
like. To some, this kind of narrative may seem chatty or not part of the documentation proper, but I would contend that it is. Somewhere in the metadata, the future users of the archived materials should read not only about the stimuli for elicitation or recording tools that were used, but also a brief characterisation of how contact was established and what the everyday research practice looked like. This kind of information will, no doubt, be highly valued when anyone in the future wants to put the documentation into perspective. It is ultimately down to the individuals involved to include (or not to include) certain items in the documentation. Since individual researchers differ in the ways in which they restrict access to a source that he or she has collected, it is important to clearly mark who made decisions of access, and preferably also why.

Digitisation and electronic archiving are themselves not neutral activities, and there are a number of different ways in which researchers (and the researched) may want to integrate these activities into their own actions. The digital promise is that the language corpora will live on for much longer than if they were not digitised, and that they will continue to grow and improve in the future. However, after some years of experience in digital archiving (reflected in this chapter), we need to realise that digital archiving not only creates new problems, but also that some of the old problems, of data access for instance, will continue. For one, we cannot guarantee that the practices that we put in place now will remain unchanged. If we did that, then the corpus would be basically dead, a body in the true sense of the word. The power of large-scale electronic collections, such as those we see growing at this moment in time, is that they go beyond the single efforts and capacity of individual researchers. The downside of this is that each researcher only has limited influence on the overall product over time. The regulatory powers are much more constrained than in a traditional single-authored data collection such as a grammar, word list or collection of texts.

Every discipline has its early adapters who embrace the new recording devices while others attempt to cling to other formats that appear to be more holistic, easier to manage and easier to adapt to local requirements. What I have suggested in this chapter is that researchers across the whole spectrum need to realise that the new technologies do not solve problems of access or contextualisation, but rather shed a particularly sharp light onto these problems, which can be a first necessary step towards addressing the issues at hand.
The new archiving effort is, to a considerable extent, being driven by the new technology of documentation and of archiving. The archivists and the technical groups involved insist that regulating access to the digital resources, as well as how to best organise these corpora, is not their responsibility, but that it remains that of the researchers in close cooperation with representatives of those who contributed to the data corpus. While this initially promises a new and wider scope for providing and sharing data with communities, it also creates some enduring problems for the researchers involved. Archiving technology is indeed changing in some fundamental ways how we generate and share knowledge. The electronic online archive is not a container that passively waits to be filled with data. Rather, it also acts as a prompt and feeds back into the research process. In this chapter, I have suggested that analysing this prompt in terms of particular features of digital records and in terms of layers of property rights facilitates orientation when participating in this complex enterprise.

References


—, ‘Property and Access Rights to the Corpus, Again’, paper presented at the DOBES (Dokumentation bedrohter Sprachen) Workshop at the Max Planck Institute for Psycholinguistics (12–13 June 2008).

1. The Archive Strikes Back


Online Sources

DOBES, CoC Code of Conduct, UID Usage Declaration, UR Usage Request, DAPR Data Access and Protection Rules, and DAA Depositor-Archivist-Agreement
<http://www.mpi.nl/DOBES/ethical_legal_aspects/>

DOBES homepage
<www.mpi.nl/DOBES>

International Association of Sound and Audiovisual Archives (IASA), Copyright & Other Intellectual Property Rights
<http://www.iasa-web.org/copyright-other-intellectual-property-rights>

Mark Liberman, Web-Based Language Documentation and Description. Legal, Ethical, and Policy Issues Concerning the Recording and Publication of Primary Language Materials

The Hans Rausing Project for Endangered Languages, Online Resources for Endangered Languages. Ethical issues
<http://www.hrelp.org/languages/resources/orel/ethical.html>

Language Archive Technology, Arbil Metadata Editor, Browser & Organizer Tool
<http://www.lat-mpi.eu/tools/arbil>