Archiving and managing sociolinguistic data: The problems of portability, access and security, and discoverability and relevance

Tyler Kendall, University of Oregon, tsk@uoregon.edu

Abstract

In recent years, sociolinguists have become increasingly focused on and more explicit about ensuring the preservation, and accessibility, of their data. This increased focus on our data has involved new lines of work explicitly on data management and, in turn, has led to important (re)considerations of the nature of sociolinguistic data and the metadata of importance for sociolinguist research. While many of the papers in this special issue focus on issues having to do with specific metadata, in this paper I consider archiving and sharing data more generally. I attend to three specific problem areas, *portability* (Bird and Simons 2003a), *access and security*, and *discoverability and relevance*, providing advice as well as some "food for thought" discussions for each.

0. Introduction

In recent years, sociolinguists have become increasingly focused on, and more explicit about, ensuring the preservation and accessibility of their data. New guidelines by grantors, such as the U.S. National Science Foundation[1] and the Canada Social Sciences and Humanities Research Council[2] are likely aiding these efforts, and new journals supporting multimedia publication

[1] See <http://www.nsf.gov/sbe/sbe_data_management_plan.jsp>.
[2] See <http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx>.

formats (e.g., the Journal of Experimental Linguistics[3] and the Journal of Linguistic Geography[4]) will surely continue to do this, but to a large extent a growing interest in the management and preservation of sociolinguistic data has come from the developments within the field itself. To quote Kendall and Van Herk (2011: 3): "The previous, dominant model of considering sociolinguistic data as too valuable to 'part with' or to share appears to be giving way to a model where sociolinguistic data is considered to be too valuable not to share."

This focus on our data has involved new lines of work explicitly on data management and, in turn, has led to important (re)considerations of the nature of sociolinguistic data and the metadata of importance for sociolinguist research. And it is these (re)considerations which are at the heart of this special issue. In this paper, I focus on three areas of data management and archiving, presenting some "food for thought" arguments about each. My comments in this paper are not meant to be taken as firm suggestions but rather as discussion points, which can hopefully fuel further conversations and developments. In a recent publication about data preservation and access in sociolinguistics (Kendall 2013a), I built on discussions of best practices in endangered language research by Bird and Simons (2003a) to argue that best practices for sociolinguistic data management, access, and preservation can and will develop from continuing conversations in the field. This special issue and the workshop it grew from represent an excellent "meeting of minds" on these issues. I use my paper here as a place to present some discussion points as we continue to explore how best to manage, preserve, and share our data.

In particular, in §1, I give some general consideration to data preservation, focusing most specifically on the time-scale of our preservation efforts and on file formats and data structure.

[3] See <http://elanguage.net/journals/jel>.

[4] See <http://journals.cambridge.org/action/displayJournal?jid=jlg>

Then, in §2, I consider questions of data sharing, discussing, in §2.1, some pros and cons of open vs. protected/limited sharing of data, and, in §2.2, the importance of ensuring that shared resources are findable by potential users and that clear specifications are given about the access options and limits for data. These three elements – what I label here respectively the *portability problem* (after Bird and Simons 2003a, see also Kendall 2013a), the *access and security problem*, and the *discoverability and relevance problem* – are far from exhaustive and I do not intend to imply that these are the most important issues in data management, archiving, and sharing (see the many of contributions to this special issue for many other important issues). I do believe that these three problem areas are generally under-addressed in the literature, including in my recent publications on the subject (Kendall 2008, 2013a, 2013b), and therefore warrant some specific attention here. I end in §3 with some closing thoughts on data management and preservation in sociolinguistics.

As a final introductory note, I should add that I use this term *problem* in this paper not entirely meaning it to have a negative sense. I view each of these "problems" as challenges facing sociolinguistic and other speech researchers, but also as representing excellent vehicles to improve our data management and general research methodologies. Thus, they are best seen as opportunities as much as problems (Kendall 2013a: 201).

1. Data preservation and the *portability problem*

Perhaps the best consideration of data preservation issues in linguistics comes in Bird and Simons' (2003a) discussion targeted at the endangered language research community (see also Kendall 2013a). In this paper Bird and Simons argue:

> Much digital language documentation and description becomes inaccessible within a decade of its creation. Linguists who have been quick to embrace new technologies, create digital materials, and publish them on the web soon find themselves in technological quicksand. Funded documentation projects are usually tied to software versions, file formats, and system configurations having a lifespan of three to five years. Once this infrastructure is no longer tended, the language documentation is quickly mired in obsolete technology… Fortunately, linguists can follow *best practices* in digital language documentation and description, greatly increasing the likelihood that their work will survive in the long term. (p. 557).

They conceptualize the problem of preservation as one of *portability*, saying that "if digital language documentation and description should transcend time, they should also be reusable in other respects: across different hardware and software platforms, across different scholarly communities (…), and across different purposes" (p. 558). This, I believe, is a compelling broader view of data preservation and, as I explore in Kendall (2013a), a useful framework for language researchers beyond the language documentation and description community. Here, I build on my recent comments in Kendall (2013a) in two areas: considering the time frames of data preservation (§1.1) and the importance of open, non-proprietary file formats (§1.2). Again, these are not meant to be an exhaustive set of topics of importance for data *portability*, but rather my goal is to dig deeper into just two areas where the recent literature (e.g. Bird and Simons 2003a, Goldman et al. 2005, Austin 2006, Kretzschmar et al. 2006, Kendall 2013a, 2013b, Schilling 2013, ch. 6) has had important things to say but which still have much room for further interrogation.

1.1. Time frames for preservation

In Kendall (2013a), I argue that researchers should consider three rough time frames when making plans for their data – the short-term, medium-term, and long-term – and that each of these time frames may involve different planning decisions and activities.

> The short-term can be thought of as the immediate future, the course of the actual research project, and one's individual interest in that data as "active" research data. The medium-term may encompass one's complete research career and/or the lifetimes of the informants in the recordings. Then, the long-term is the unforeseeable future – what use can future scholars gain from the data as a part of the historical record of a language variety or a community? (Kendall 2013a: 199)

A challenge for researchers, given all of the other demands on our time and our more immediate desires (such as the actual research outcomes), is that planning for each of these three time-frames is best done early in the research process, even – or perhaps especially – planning for the long-term storage, preservation, and possible sharing of the data. This is both for ethical and compliance reasons (see e.g. Warner this issue) and for practical planning purposes. The biggest difficulty in sharing data is not technical but is more simply about having the proper permissions and rights in place to share the data, and permissions, both from the participants in the research and the relevant ethics or human subjects board, are best obtained up front, before gathering the data. Meanwhile, while fully annotating our data and fleshing out its metadata elements (see e.g. the papers in the second half of this issue) can feel like time away from our actual research interests, committing our knowledge "to paper" so to speak before our memory fades ultimately saves a great deal of time and results in more accurate metadata than if these steps are undertaken late in the research process or only after the immediate research goals have been met. This may seem like obvious advice but I believe it is important to stress and to remember. I admit I have often skipped important steps in organizing and marking up my data in order to "jump ahead" to my actual research goals and have ended up cursing myself for having made the work that much harder. In sum, at the start of every research project *give full consideration to short-term, medium-term, and long-term plans for the data*.

1.2. The importance of open, non-proprietary file formats

Bird and Simons (2003a) offer a number of pieces of concrete advice, some of which can boil down to *don't trust electronic formats and media for long-term storage and preservation*. Formats change and technologies and software (and companies) come and go. A common – and good – recommendation in the literature (cf. Simons and Bird 2003a, Austin 2006, Farrar and Lewis 2007) is to follow emerging standards, like the use of particular instantiations of XML, such as the TEI guidelines (Burnard and Bauman 2007). Following standards in developing your metadata has numerous benefits. Primarily it ensures that the way you code your data, for yourself and for other potential users, matches common practices – ideally best practices – in the field, and thus builds on the work of others. In addition to helping you account for important metadata you may not be considering for your own research but which might still be helpful, it also ensures that potential users of your data can readily interpret your data. What does, for instance, "middle class" mean for your particular dataset? Or, even more opaque, how do you interpret the code "S07M2C3" years from now if such a coding system was used in file-naming or as header information?

While it is good advice to follow standards in the field – and I certainly do not mean to counter that advice here – a problem with emerging standards to non-experts is that they seem to be constantly "emerging" and therefore present a moving target. Engaging fully in the academic discourse on standards can bring scholars primarily interested in working on their substantive research far afield from their main projects. For example, the Text Encoding Initiative hosts an academic journal, an annual meeting, special interest groups and so forth, and posts updates to its (1,500+ page) guideline document several times a year.[5] We clearly all cannot become experts on this stuff.

While it is important to be aware of accepted standards for things like metadata and some of the flux of emerging standards will certainly dissipate as these standards are refined and more widely accepted, one simpler piece of more immediate advice is to document your own coding decisions, including not only what criteria are used to determine categories (again, e.g., "middle class" but also linguistic categories like "habitual" or "non-apical" in the coding of linguistic variables) but also how these are actually stored and encoded in your files. And even more importantly, *never* rely on formats that are not readable as plain text. XML documents, Praat TextGrids, ELAN .eaf files, and many other software file formats can be read into simple text editors or computer programming scripts and thus can be easily parsed or modified, without the need for the original software. Other formats – MS Word and MS Excel files come most to mind – are stored as binary files and cannot be read without the original software or some other specialized reader (like Google Docs or OpenOffice). As Bird and Simons (2003a) cogently argue, there is no telling which of these software packages will survive the test of time and which formats will remain readable into the future. At the same time, use of readable file formats is not

[5] See <http://tei-c.org/>. The TEI is, admittedly, an extreme example; not all standards are nearly so complex.

enough unless the files are extremely simple, open and documented. Information also needs to be stored that explains the details of the formatting. How do future users (even your-future-self) read the file and interpret its contents?

Let us take Praat as an example. For those unfamiliar, Praat is a popular phonetics analysis software package (Boersma and Weenink 2001-2013[6]). In addition to being used for acoustic analysis, Praat includes a suite of other features, including e.g. features for articulatory and acoustic speech synthesis, Optimality-Theoretic and Harmonic-Grammar learning/analysis, and various kinds of audio annotation. In particular, Praat is a useful annotation and transcription package, as its TextGrid object allows researchers to generate diverse tiers of information, from orthographic text transcription to quantitative or qualitative coding, all finely time-aligned to the source audio (see, e.g., Kendall 2007). Praat allows TextGrids to be saved into several different kinds of formats, including what Praat's file saving menu terms a "text file" (the typical/default format), a "short text file", and a "chronological text file". However, and regardless of the file extension (.TextGrid, .TG, .txt, …), each of these is underlyingly just a text file with specific formatting. Figures 1 - 4 display screenshots of different versions of the same TextGrid transcript. Figure 1 displays the Praat editor window where we see Praat's interface to the audio and TextGrid together. Figure 2 displays the beginning of the TextGrid saved in its basic format ("Save as text file…"[7]) in Praat version 5.3.05 and opened in the TextEdit application on Mac OS 10.7. Figure 3 displays the same TextGrid saved using "Save as chronological text file…" using the same version of Praat and viewed in TextEdit. Figure 4 displays the same "chronological text" file as saved from Praat version 5.2.17. Note that each of

[6] See <http://praat.org>.

[7] Note, in older Praat versions these save menu options where termed "Write to…" rather than "Save as…".

these is different, including the two chronological text files saved from the two different versions of Praat (Figures 3 and 4). The main point here is that each of these versions can be read as plain text and each of these versions is more or less interpretable. (Comparing the four figures you should be able to reconstruct what each piece means.) But, if we were to rely on a script (such as a script in the R language; R Development Team 2013) to parse a number of TextGrid files for a particular purpose, we would need to know both the format of the files and the version of Praat used to save them. If we develop a collection of TextGrid transcripts over time, using different versions of Praat, there is no guarantee that all of our files will be exactly identical. In sum, *it pays to get to know your file formats and to document them for all of your files.*

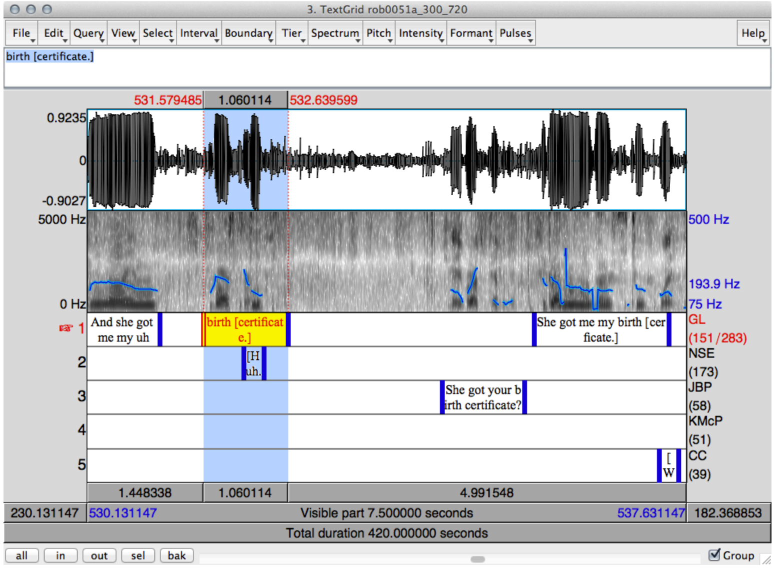
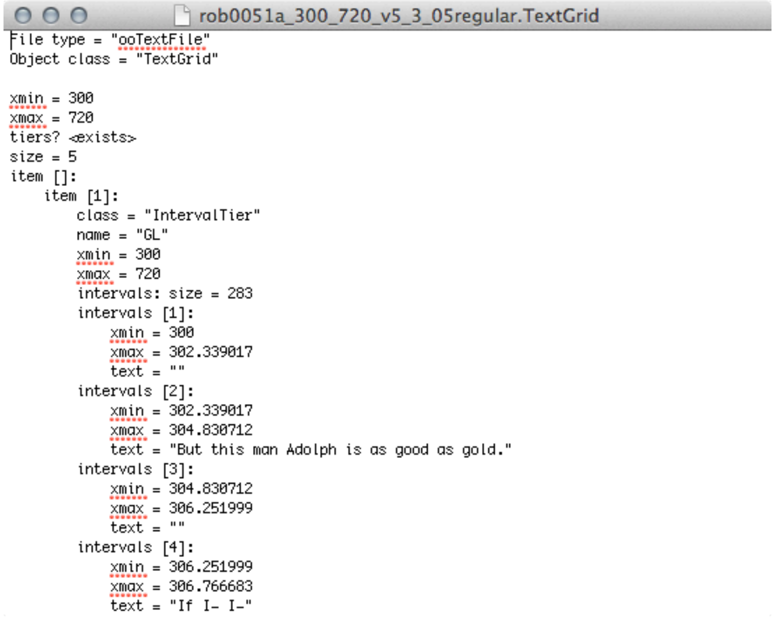


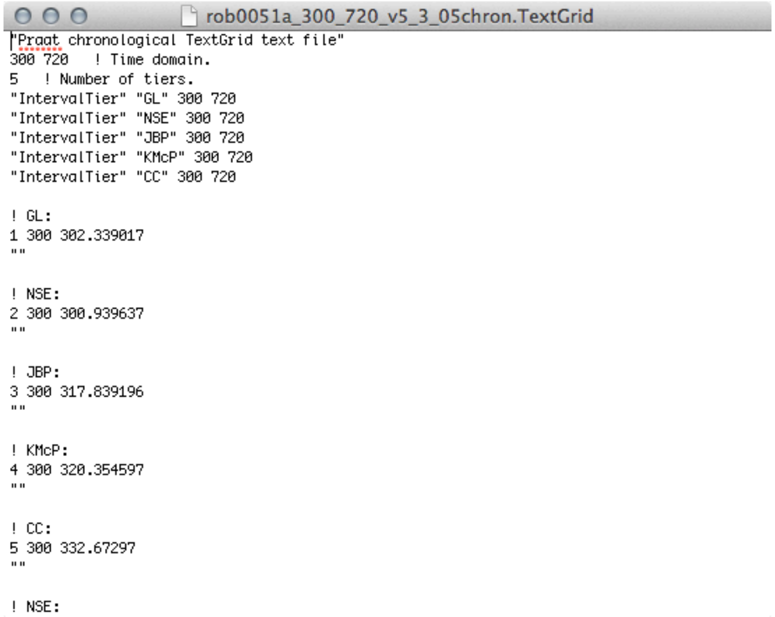
Figure 1. Praat Editor Window showing audio and TextGrid transcript



rob0051a_300_720_v5_3_05regular.TextGrid

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 300
xmax = 720
tiers? <exists>
size = 5
item []:
    item [1]:
        class = "IntervalTier"
        name = "GL"
        xmin = 300
        xmax = 720
        intervals: size = 283
        intervals [1]:
            xmin = 300
            xmax = 302.339017
            text = ""
        intervals [2]:
            xmin = 302.339017
            xmax = 304.830712
            text = "But this man Adolph is as good as gold."
        intervals [3]:
            xmin = 304.830712
            xmax = 306.251999
            text = ""
        intervals [4]:
            xmin = 306.251999
            xmax = 306.766683
            text = "If I- I-"
```

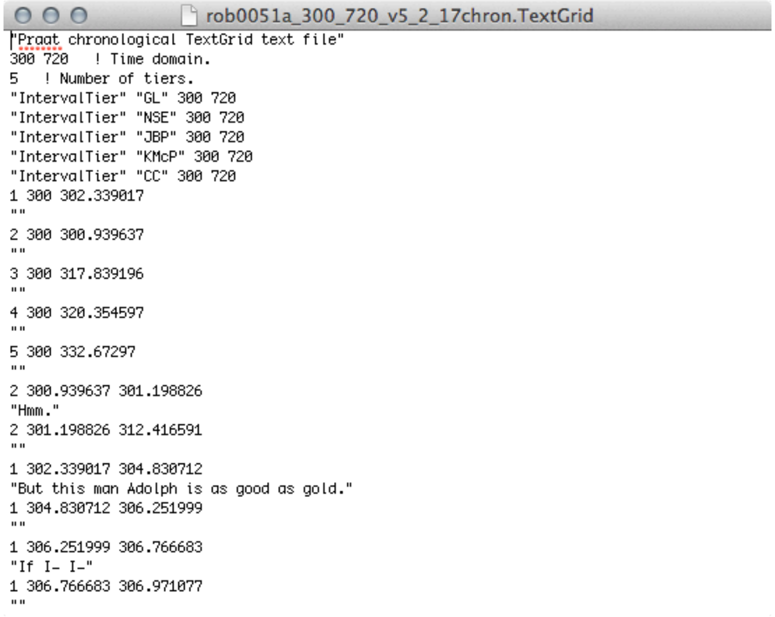
Figure 2. Praat (ver. 5.3.05) TextGrid transcript opened in TextEdit (Mac OS 10.7)



rob0051a_300_720_v5_3_05chron.TextGrid

```
"Praat chronological TextGrid text file"
300 720   ! Time domain.
5   ! Number of tiers.
"IntervalTier" "GL" 300 720
"IntervalTier" "NSE" 300 720
"IntervalTier" "JBP" 300 720
"IntervalTier" "KMcP" 300 720
"IntervalTier" "CC" 300 720

! GL:
1 300 302.339017
""

! NSE:
2 300 300.939637
""

! JBP:
3 300 317.839196
""

! KMcP:
4 300 320.354597
""

! CC:
5 300 332.67297
""

! NSE:
```

Figure 3. Praat (ver. 5.3.05) chronological TextGrid opened in TextEdit (Mac OS 10.7)



```
"Praat chronological TextGrid text file"
300 720   ! Time domain.
5   ! Number of tiers.
"IntervalTier" "GL" 300 720
"IntervalTier" "NSE" 300 720
"IntervalTier" "JBP" 300 720
"IntervalTier" "KMcP" 300 720
"IntervalTier" "CC" 300 720
1 300 302.339017
""
2 300 300.939637
""
3 300 317.839196
""
4 300 320.354597
""
5 300 332.67297
""
2 300.939637 301.198826
"Hmm."
2 301.198826 312.416591
""
1 302.339017 304.830712
"But this man Adolph is as good as gold."
1 304.830712 306.251999
""
1 306.251999 306.766683
"If I- I-"
1 306.766683 306.971077
""
```

Figure 4. Praat (ver. 5.2.17) chronological TextGrid opened in TextEdit (Mac OS 10.7)

While it can be frustrating that file formats change and that software programs implement features differently in different versions, the most important point to take away here is that this whole Praat exercise is possible only because we can readily open the Praat files and examine their contents using a plain text reader.

2. Data sharing

Assuming we intend to make our data accessible to individuals outside our immediate research/colleague group, a number of important issues come up having to do with the question of exactly how we actually do this. The most obvious approach given the current status of the World Wide Web as a ubiquitous global resource is to post our data on a website. However, sharing our data over the web is often not as simple as just posting it to a website. Web pages can be quite ephemeral and even simple, static HTML-based pages can cease to work over time (for instance, as an institution's information technology division updates or otherwise changes the way that web services are offered). Further, we may in fact not want our data accessible to everyone on the web, in which case a number of important considerations need to be made about how exactly to share the data without sharing it wholesale. I address this, the *access and security problem*, in §2.1. Simultaneously, we also want to make sure that our particular data are findable by the right users. It turns out there is a lot of "stuff" on the web and lots of times we will find a "corpus of X" on the web thinking that it might be exactly what we have always needed for our research only to discover that it is in fact not available to us or not at all what it looked like from its name or description. Making sure that relevant users actually find your data and know that it is relevant and useable to them is an important task (Bird and Simons 2003a, 2003b). I have termed this here the *discoverability and relevance problem* and address it in §2.2.

2.1 Access and security

It seems to me that very few sociolinguistic data collections are necessarily the kinds of resources that need to be posted on the open web, downloadable by everyone at the click of the mouse with no delay or restrictions. Of course, making one's data available in this way is wonderful and, provided the permissions and rights are in place (see Warner, this issue), posting resources on the open web is probably the easiest way to share data from a technical perspective (although see Kendall 2013a and Bird and Simons 2003a for discussions of how posting files to the web does not ensure long-term preservation or access; also see Goldman et al. 2005 on the ephemeral nature of web-based audio). But, password protecting or establishing a licensing agreement seems to me perfectly reasonable – especially for data that you are still using (the short-term) or potentially may still put to new uses (the medium-term). Putting your data online with no clear restrictions or limits essentially gives permission, whether explicitly or not, that any user can do whatever they wish with it. It can be quite surprising – in a bad way – to, for instance, attend a conference to find that someone is presenting research using your data on a topic you want to pursue yourself or to discover an excerpt from an interview in your data used in a news report or in a blog post. Whether or not the data are put on the open web, establishing some restrictions or at least thinking through the ramifications of not doing so, is a valuable safeguard. Once data are posted online as freely downloadable their availability to the public cannot be undone; even taking the data off the web cannot ensure that people did not download the data when it was available. It is not practical, or for that matter fair to bona fide users, to assume that you can later rescind "public access" once you have given it over the open web.

At the same time, using some sort of password protection system, so that users can access the data only after receiving a password has the benefit of letting you negotiate terms with potential users. I personally advocate for the widest possible sharing of data, but if you created the data you certainly have the rights to know, and limit if you really desire, what potential users are going to do with your data. The key here, I think, relates to the next issue – discoverability and ensuring that potential users find your data and understand how they can use it, and what their rights are.

2.2 Discoverability and relevance

A number of researchers have created quite impressive corpora in the history of sociolinguistics (e.g., Poplack's Ottawa-Hull French Project (cf. Poplack 1989), Tagliamonte's York English Corpus and Toronto English Archive of Spoken Materials (cf. Tagliamonte 2012), and Labov's Philadelphia Neighborhood Corpus (cf. Labov, Rosenfelder, and Fruehwald 2013), to name just a few), but it seems to be becoming fairly common practice in sociolinguistics to refer to even our small data collections as “corpora” or “databases” using proper names to describe them in our research (e.g., the fictitious *Corpus of Oregonian English*). I will argue that quite often these “corpora” are not corpora in the sense that corpus linguists use the term – publically accessible, large, machine-readable resources (see McEnery and Wilson 2001, Kendall 2011, 2013b). This is fine – of course – as terms are just terms and different (sub-)disciplines can use terms differently, but the practice of naming small datasets can be misleading, especially if those data are not shareable or accessible by others.

If data are shareable they should be discoverable in relatively straightforward ways. By this, I mean they should be findable on the web and have clear instructions about how one can access them. For many years I have discussed the Sociolinguistic Archive and Analysis Project (SLAAP[8]; cf. Kendall 2007, 2008, Kendall and Bradlow 2011) as an example of a speech data management system and many of the researchers whose data reside in SLAAP have been, in my opinion, quite gracious in sharing these data with others. However, for most of SLAAP’s existence, there was no easy way for potential users to find out exactly what data collections were in SLAAP short of asking and, as a result, I often received requests for data nothing like those in SLAAP. And I am sure that I frustrated many researchers by not making it clear what materials actually were in SLAAP. (Further, I, as the archive administrator, spent a lot of my time fielding questions that could have been resolved without my active attention had better information been in place about the accessible data.) Beginning in 2012, SLAAP joined the network of language repositories in the Online Language Archives Community (OLAC[9]; cf. Simons and Bird 2003) and now describes most of the collections available in SLAAP in the searchable OLAC database. This way, we hope, potential users can peruse the collections in

[8] See <http://slaap.lib.ncsu.edu/>.
[9] See <http://www.language-archives.org/>.

SLAAP easily and will also discover that resources are available if, for instance, searching the web for “well-known” sociolinguistic studies.

3. In closing

My own work developing and hosting the SLAAP sociolinguistic data repository over the years has put me face-to-face with a number of issues in data management, preservation, and sharing. It has also made it clear to me that there are no simple answers to many of the important questions when it comes to these issues. If there were, we would not need special issues like this one, or organizations devoted to data sharing and interoperability, or centralized archives. But it is exactly for this reason that we must continue to explore and discuss best practices for archiving and sharing spoken language data. One thing is for sure: our data are too valuable to lose or to let fade away over time. Sociolinguistic recordings are important records which capture particular cultural positions at particular times and which can preserve specific moments in the history of language varieties. They also represent valuable potential resources for large-scale aggregation and analysis (Coleman et al. 2011). By improving our data management, sharing, and preservation practices, sociolinguists can enhance the impact of our research and our contributions to human knowledge about language and society.

References

Austin, Peter K. 2006. Data and language documentation. Essentials of language documentation, ed. by Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 87-112. Berlin: Mouton de Gruyter.

Bird, Steven and Gary Simons. 2003a. Seven dimensions of portability for language documentation and description. Language 79(3). 557-82.

Bird, Steven and Gary Simons. 2003b. Extending Dublic Core metadata to support the description and discovery of language resources. Computers and the Humanities 37(4). 375-88.

Boersma, Paul and David Weenink. 2001-2013. Praat: Doing phonetics by computer. Amsterdam: Phonetic Sciences, University of Amsterdam. [Computer program]

Burnard, Lou and Syd Bauman. 2007. P5: Guidelines for electronic text encoding and interchange. TEI Consortium. <http://tei-c.org/>

Coleman, John, Mark Liberman, Greg Kochanski, Lou Burnard, and Jiahong Yuan. 2011. Mining a year of speech. Paper presented at New Tools and Methods for Very-Large-Scale Phonetics Research. University of Pennsylvania: Philadelphia, PA.

Farrar, Scott and William D. Lewis. 2007. The GOLD community of practice: An infrastructure for linguistic data on the Web. Language Resources and Evaluation 41(1). 45-60.

Goldman, Jerry, Steve Renals, Steven Bird, Franciska de Jong, Marcello Federico, Carl Fleischhauer, Mark Kornbluh, Lori Lamel, Douglas W. Oard, Claire Stewart, and Richard Wright. 2005. Accessing the spoken word. International Journal on Digital Libraries 5(4). 287-98.

Kendall, Tyler. 2007. Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. Penn Working Papers in Linguistics 13(2). 15-26.

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. Language and Linguistics Compass 2(2). 332-51.

Kendall, Tyler. 2011. Corpora and from a sociolinguistic perspective (Corpora sob uma perspectiva sociolinguística). Corpus studies: Future directions, special issue of Revista Brasileira de Linguística Aplicada, ed. by Stefan Th. Gries, 11(2). 361-89.

Kendall, Tyler. 2013a. Data preservation and access. Data collection in sociolinguistics: methods and applications, ed. by Christine Mallinson, Becky Childs, and Gerard Van Herk, 195-205. New York: Routledge.

Kendall, Tyler. 2013b. Data in the study of variation and change. The Handbook of Language Variation and Change, 2nd edition, ed. by J. K. Chambers and Natalie Schilling, 38-56. Malden, MA/Oxford: Wiley-Blackwell.

Kendall, Tyler and Ann R. Bradlow. 2011. Mobilizing smaller datasets for large-scale phonetic analysis: web-databases and semi- automatic analyses. Paper presented at New Tools and Methods for Very-Large-Scale Phonetics Research. University of Pennsylvania: Philadelphia, PA.

Kendall, Tyler and Gerard Van Herk. 2011. Corpus linguistics and sociolinguistic inquiry: Introduction to special issue. Corpus Linguistics and Linguistic Theory 7(1). 1-6.

Kretzschmar, William Jr., Jean Anderson, Joan Beal, Karen Corrigan, Lisa Lena Opas-Hänninen, and Bartlomiej Plichta. 2006. Collaboration on corpora for regional and social analysis. Journal of English Linguistics 34(3). 172-205.

Labov, William, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in Philadelphia: Linear incrementation, reversal, and reanalysis. Language 89(1). 30-65.

McEnery, Tony and Andrew Wilson. 2001. Corpus Linguistics, 2nd edition. Edinburgh: Edinburgh University Press.

Poplack, Shana. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French Project. Language Change and Variation, ed. by Ralph Fasold and Deborah Schiffrin, 411–51. Amsterdam: John Benjamins.

R Development Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. [Computer program]

Schilling, Natalie. 2013. Sociolinguistic Fieldwork. Cambridge: Cambridge University Press.

Simons, Gary and Steven Bird. 2003. The Open Language Archives Community: An infrastructure for distributed archiving of language resources. Literary and Linguistic Computing 18. 117–28.

Tagliamonte, Sali. 2012. Variationist Sociolinguistics: Change, Observation, Interpretation. Malden, MA/Oxford: Wiley-Blackwell.