Toward open data policies in phonetics: What we can gain and how we can avoid pitfalls

Marc Garellek, Matthew Gordon, James Kirby, Wai-Sum Lee, Alexis Michaud, Christine Mooshammer, Oliver Niebuhr, Daniel Recasens, Timo B Roettger, Adrian Simpson, et al.

Journal of Speech Sciences (9): 03-16. 2020.
Available at: http://revistas.iel.unicamp.br/joss

LETTER TO THE EDITOR: TOWARD OPEN DATA POLICIES IN PHONETICS: WHAT WE CAN GAIN AND HOW WE CAN AVOID PITFALLS

GARELLEK, Marc[1]
GORDON, Matthew[2]
KIRBY, James[3]
LEE, Wai-Sum[4]
MICHAUD, Alexis[5]*
MOOSHAMMER, Christine[6, 7]
NIEBUHR, Oliver[8]
RECASENS, Daniel[9]
ROETTGER, Timo B.[10]
SIMPSON, Adrian[11]
YU, Kristine M.[12]

[1]Department of Linguistics, University of California, San Diego
[2]Department of Linguistics, University of California, Santa Barbara
[3]Department of Linguistics and English Language, University of Edinburgh
[4]Department of Linguistics and Translation, City University of Hong Kong
[5]Langues et Civilisations à Tradition Orale, CNRS – Sorbonne Nouvelle
[6]Sprach- und literaturwissenschaftliche Fakultät, Humboldt-Universität zu Berlin
[7]Haskins Laboratories
[8]Mads Clausen Institute Sønderborg, University of Southern Denmark
[9]Departament de Filologia Catalana, Universitat Autònoma de Barcelona
[10]Institute of Cognitive Science, Universität Osnabrück
[11]Institut für germanistische Sprachwissenschaft, Friedrich-Schiller-Universität Jena
[12]Department of Linguistics, University of Massachusetts Amherst

Abstract: *It is not yet standard practice in phonetics to provide access to audio files along with submissions to journals. This is paradoxical in view of the importance of data for phonetic research: from audio signals to the whole range of data acquired in phonetic experiments. The phonetic sciences stand to gain greatly from data availability: what is at stake is no less than reproducibility and cumulative progress. We will argue that a collective turn to Open Science holds great promise for phonetics. First, simple reflections on why access to primary data matters are recapitulated and proposed as a basis for consensus. Next, possible drawbacks of data availability are addressed. Finally, we argue that data curation and archiving are to be recognized as part of the same activity that results in the publication of research papers, rather than attempting to build a parallel system to incentivize data archiving by itself.*

Keywords: experimental phonetics; phonetic sciences; research data; data curation; data conservation; open access; open archives; open science.

*Corresponding author: alexis.michaud@cnrs.fr

Introduction: Phonetics and Open Science

The brimming agendas of researchers in the phonetic sciences have in recent years received the unsolicited addition of increasingly complex and time-consuming tasks related to the protection of human subjects' personal data. Database creation and management are associated with administrative hurdles that sometimes seem to reflect institutions' concerns about liability (protecting themselves from legal lawsuits) more closely than the real concerns of the human subjects and of the researchers. Overall, researchers are made to feel that they have a legal (and moral) obligation to restrain access to data. This has the negative side effect of discouraging data sharing: researchers who are not 100% confident that it is legally and ethically fine to allow access to data that they collected may go for the cautious option of keeping data private. Data curation and archiving are also impacted: if there are strong reasons why data cannot be shared, why take the trouble to prepare data sets for future reuse by others?

Due in part to these legal concerns and in part to inertia (continuing editorial habits that predate the time when data sharing was a technical possibility), it is not yet standard practice in phonetics to provide access to audio files along with submissions to journals. This is paradoxical in view of the importance of primary data in our field (emphasized e.g. by 1). *Primary data* includes all the data that we collect, from audio signals to the whole range of data acquired in phonetic/phonological experiments. The emphasis laid on legal (and moral) obligations to restrain access to data threatens to override the strong reasons for data sharing.

Given this context, it appears useful to state the obvious: that the phonetic sciences stand to gain greatly from data availability and citability (2). Data sharing is a crucial and necessary part of responsible conduct in research. Research data need to be curated, archived, and made *as open as possible, as closed as necessary*. As in other fields of science, what is at stake is no less than reproducibility and cumulative progress.

> "The American National Academy of Sciences Committee on Science, Engineering and Public Policy says unambiguously that researchers have a 'fundamental obligation' to keep quality records of their research, and that once it is published, that other researchers 'must have access to the data and research materials'. This responsibility flows from several sources: our epistemic responsibility to substantiate our claims with evidence, our responsibility to the community of scientists from whom we obtained most of our knowledge, and our responsibility to society at large that supports our research." (3)

But the field of phonetics and phonology lacks a unified set of guidelines on how to achieve this goal. Efforts to curate and archive data sets currently depend on the goodwill of individuals. The present document argues that a collective turn to make data curation and archiving mandatory will be highly beneficial for the phonetic sciences. The broader background to this argument is a general reflection about Open Science (4,5). Open and reproducible research practices in phonetic sciences are advocated in various publications: for instance, (6) argues for transparent sharing of all aspects of a research projects including materials, raw data, data tables, and scripts. By contrast, the present argument focuses specifically on the issue of access to primary data.

Copying a *caveat* from an article about another field of research, we note that the present argument is made "by insiders offering a critical introspective look, not as sniping outsiders" (7). Our hope is to facilitate a base level of common understanding so that the field can deal with these core issues actively and manage ongoing transitions tactfully, rather than passively

letting changes happen around us and belatedly realizing that data have evaporated and many research articles in phonetics play to our ears "ditties of no tone", like the silent Grecian urn in John Keats's *Ode*.[i]

First, simple reflections on *why access to primary data matters* are recapitulated and proposed as a basis for consensus. Next, possible drawbacks of data availability are addressed. Finally, a set of recommendations is proposed.

1 Why Access to Primary Data Matters

The importance of primary data (audio files, as well as data collected through the entire range of exploratory techniques used in the phonetic sciences) is self-evident, but given the current situation of disconnect between publications and data, it many nonetheless be useful to spell out some fundamental (and, we think, uncontroversial) points. Emphasis is placed here on benefits for individual researchers, as well as for the field (see 8). Publicly sharing data has several advantages, including a substantial citation benefit (9) and increased opportunities for funding, jobs, and collaborations (10). Additionally, storing data on permanently accessible repositories protects against data loss and link rot. Beyond benefits for the individual researcher, open data and materials allow other researchers to reproduce, replicate or extend our findings without duplicating efforts, saving valuable resources and allowing for a more rapid advancement of knowledge (11).

1.1 Access to Data is Useful for Reviewing

"[P]ublications are arguments made by authors, and data are the evidence used to support the arguments" (12). Without a sample of the recordings on which the analysis is based, we do not think that it is feasible to provide a reliable and fair review of a manuscript. Appraising the relevance and validity of the statistical treatment of the data is one (nontrivial) thing: it amounts to checking a manuscript's consistency as a self-contained piece of writing. But beyond this in-depth check, peer review involves evaluating arguments that pertain to the data. Lack of access to the primary data places obvious limitations on this process. This is particularly relevant for quantitative assessments in phonetics, as speech analysis allows for many degrees of freedom in measurement and analysis, effectively inviting human bias during data interpretation (6,13).

When a manuscript is about a language with which one is familiar, the sounds at issue may seem self-evident. But this sense of transparency can be misleading, in view of the phonetic diversity of one and the same language (14). When one is not familiar with the language under study, the lack of audio data is an even more pressing issue.

Access to full data sets and tools is obviously most appealing, as it makes it possible to reproduce an experiment in full.

1.2 Access to Data Increases the Value of an Article for its Readers

Readers also value access to the data. For instance, readers who are unsure how aryepiglottic sounds differ from pharyngeal ones will be relieved to have access to audio recordings when reading an article on this topic. But even for topics (and languages) with which one is thoroughly familiar, it is no less important for readers than for reviewers to be able to listen to audio examples as a means to get a feel for the phenomena at issue.

Disagreement across studies is not uncommon for many reasons, which include differences in choices made during data collection, differences in target dialect, in consultants' understanding of the tasks entrusted to them, and other factors which can be grasped by

examining the data sets of the various studies. Sophisticated exploratory techniques and statistical data analysis do not make the ear obsolete as a research tool. We can get the best of both worlds: combining instrumental, quantitative and theoretical analyses with insights based on contact with primary data. Access to data can shed light on seemingly contradictory findings. It yields a deeper understanding of the facts, placing the various studies in a perspective where they supplement one another. It can also suggest new avenues for exploration, thereby stimulating new research. Nowadays, it is not so frequent for authors to receive correspondence from readers (colleagues, students, or members of a wider audience). Access to data has potential to stimulate scientific exchanges.

Increasingly, online versions of linguistic papers have in-text links to the audio of example sentences, and to various other materials. Online data hosting for the short term may seem easy and inexpensive, but URLs can be flimsy. State-of-the-art data curation and archiving is clearly the way to go, so that links to data do not break over time. Primary data hosted in archives are taken care of for the long term. Stable identifiers include types that we phoneticians seldom manipulate, and about which we may not want to learn a great deal of technical detail, but which are of great value in the mid and long run: the California Digital Library's ARK (Archival Resource Key), for instance. Digital Object Identifiers (DOIs) are currently very popular and can offer one-click links to data, facilitating navigation between publications and data (15).

1.3 Access to Data Allows for Cumulative Progress

Evidence from a single study is limited. To assess the robustness of a speech phenomenon, it is important to gather evidence across several similar studies and re-evaluate it (16). Having access to data would allow authors to critically evaluate published claims by running alternative measurements and analyses on existing data. This is tremendously important as many aspects of research are subjective. For example, 29 teams involving 61 analysts used the same data set to answer one question in Silberzahn et al. (17). Two thirds found a statistically significant effect, one third did not, emphasizing that analytical flexibility can have consequential impact on data interpretation. Moreover, having access to data allows us to synthesize evidence using meta-analysis. Meta-analyses are quantitative assessments of evidence across multiple studies (see 16 for a phonetic example). Access to primary data allows us to carry out a quantitative evaluation of our knowledge landscape, allowing for a swifter accumulation of knowledge.

An example where researchers making their phonetic data public allowed follow-up work is re-use of the Shilluk alignment data collected and annotated by Bert Remijsen and Otto Gwado Ayoker. The data consist of Shilluk utterances from ten speakers, obtained through controlled elicitation. The data were collected in order to investigate the phonological contrast between Low vs. Early High Fall vs. Late High Fall vs. High in Shilluk. This dataset, which forms the basis of a study of contrastive tonal alignment in falling contours in Shilluk (18), was made available through Edinburgh DataShare, a digital repository of research data produced at the University of Edinburgh (https://datashare.is.ed.ac.uk/handle/10283/633). Data availability made possible a reanalysis by a different team of authors, who proposed that the tonal contrast previously described as a typologically unusual distinction between two falling contours of identical shape and magnitude, differing only in the timing of the fall within the syllable, in fact involves distinctions in both the f_0 timing and scaling domains (19).

1.4 Making Data Duration Mandatory for Publication can Curb Data Evaporation

Put simply, a decisive reason to deposit our data in an archive is so that we don't lose them. A negative consequence of the lack of association of publications to data is that most data sit on researchers' hard drives, without the basic curation that would be necessary for these data sets to be intelligible to others than the team who recorded them. Loss of data is a troubling trend in contemporary science as a whole. Accordingly, advocacy for a change in practices comes from various fields of science. Reproducible research involves the capacity of other researchers (who have not conducted the original study) to repeat the analysis that is presented in a published study. Reproducibility necessitates that both the primary data and the analysis code are made available to the community, if this is possible (8).

The final qualification, "if this is possible", offers an opportunity to turn to a discussion of causes of the current situation of data unavailability.

2 Obstacles to Curation and Archiving of Primary Data

Since there are so many pressing reasons to share data, and so many benefits to data sharing, how come Open Science practices are not yet standard in our field?

There is a risk that the suggestion to open up research data will be a flag-raiser. Debates about topics of Open Science, such as open access to publications and data, require careful framing.

> "... while Open Access to publications originated from a grassroots movement born in scholarly circles and academic libraries, policymakers and research funders play a new prescribing role in the area of (open) scholarly practices. This adds new stakeholders who introduce topics and arguments relating to career incentives, research evaluation and business models for publicly funded research (...), surfacing old and new tensions. (...) this highlights (...) the need for better-informed debates." (20)

Data sharing can place an individual scientist at risk through (i) exposure of data to the prejudiced scrutiny of competitors or detractors, (ii) risk of compromising confidentiality of human subjects, (iii) loss of credit or opportunity, and (iv) expense of time and resources to meet requests for archiving or sharing of data.[ii] These concerns are addressed successively below.

2.1 Concerns of Criticism

A first concern is "exposure of data to the prejudiced scrutiny of competitors or detractors".[iii] Accessible data could earn the authors criticism: the value of the data may be called into question, reflecting negatively on the research article and its authors. One may worry that data that are made available on the open internet will become the object of unfair criticism. Once one is confident about painstakingly obtained experimental results, and when these results have been written up and accepted for publication by a journal after rounds of reviewing, it is tempting to consider that the results have been validated for good, and that there is no need for further re-examination of the primary data. But this is of course a slippery slope: the number of quantitative studies that fail to replicate is a case in point (e.g. 21). Humans are prone to cognitive biases, as abundantly documented in works of literature.

> "That looked good. Yes, that looked very good. In fact it went on looking better and better, straight along—until by-and-by it grew into positive proof. And then Richards put the matter at once out of his mind, for he had a private instinct that a proof once established is better left so." (Mark Twain, *The Man That Corrupted Hadleyburg*, 1899)

Craving coherence, we may see patterns in randomness (22). We weigh evidence in favour of our preconceptions more strongly than evidence that challenges our entrenched beliefs (23,24). We perceive events as plausible and predictable after they have occurred (25). Data sharing is a healthy way to counterbalance these biases.

2.2 Concerns of Privacy and Research Ethics

Another concern is that opening up data is not possible because it would be unlawful or unethical. Both arguments act as powerful deterrents, because no sane researcher would want to break the law or to act contrary to research ethics. But as far as legal issues are concerned, there is nothing illegal about sharing research data so long as the participants in the experiments agree to it, for instance placing the data under Creative Commons licenses.[iv]

Ethical issues are a different topic, which needs to be handled sensitively, with due attention to the huge diversity of situations concerned. The ethical issues are important and need to be acknowledged up front. Respect for the language consultants' wishes, choices and cultural concerns about the recorded data is an integral part of the respect that we owe them. But given that we are already (or certainly should be) getting a record from each participant of how they are agreeing to let their data be used, for those participants that agree, what is the argument against making the recordings publicly available? If the recordings in question are sensitive materials, or stories about a family or community, we can see participants not wanting them to be available. But the type of data collected for phonetic studies is often much less sensitive than, say, data from sociological or ethnological investigations – or than personal information posted on social media by users under terms and conditions that are by no means transparent, and that are not shaped by considerations of public interest or respect of privacy.

Individual researchers submitting protocols for approval by Institutional Review Boards are left on their own to deal with requirements that are not tailored to the sorts of data commonly handled in the phonetic sciences. Thus, in the United States, a major concern at present is identifiability. In clinical research, disclosing patients' medical files is professional misconduct. Some materials recorded for clinical phonetic research, once connected to the identities of the human subjects, would disclose confidential medical information. Since an individual's voice is unique, audio files contain de-anonymizing information. In fact, in view of the rapid progress of speech processing software, it is not at all unlikely that people will be uniquely identifiable through a small sample of audio in the foreseeable future. It is thus impossible to guarantee that someone's identity will not be uniquely identifiable in future through their voice, even if the data that they recorded is "de-identified" by using identifiers that keep the consultants' names anonymous. In the case of hereditary diseases, disclosure of information could reflect on whole generations of relatives. Obviously, stringent data management rules should apply for clinical phonetics. Not so many data sets used in phonetic research are that sensitive, however. No, we can never guarantee that the person will not be identified via their voice, but identification need not always be seen as a problem by itself. Discussion with consultants about phonetic data being made public (audio, ultrasound images of their tongue movements, or EMA data or such) may even bring out some consultants' preference for their identity to be made public when a data set is released, as a recognition of their contribution to a scientific enterprise. Anonymizing data, or keeping the data to oneself,

may not be as ethical as it seems. One needs to ask to what extent the real motivation is to protect the consultants, and to what extent the motivation is self-protection (to be safeguarded against potential trouble). Lack of data conservation and of access to the data could also be seen as unethical behaviour, to the extent that it deprives the community of a useful resource.

A statement from phoneticians/phonologists as a field (perhaps through learned societies such as the International Phonetic Association, the Association for Laboratory Phonology, the International Speech Communication Association and the Acoustical Society of America) would play an important role here, by offering an articulate point of view. Reference materials from within our community would facilitate dialogue with Institutional Review Boards and Ethics Committees, in the interest of true ethics and good science.

2.3 Concerns of Data Appropriation

In the field of experimental phonetics, we are accustomed to keeping our research data private. Because data acquisition takes lots of effort and resources – often more than initially anticipated (see e.g. 27) –, giving data away for free tends to be seen as giving a helping hand to competitors who can then publish faster, effectively scooping the generous donors and placing them at a disadvantage in a context where there is strong competition for funding and positions. There are some particularly sensitive research areas, such as topical issues in the speech processing industry. Expressive speech and voice charisma are a case in point. The authors of the patent-pending automatic public-speaker assessment system PASCAL (28) recorded a multi-language speaker database (about 400 speakers at the moment). In connection with the patent and the financial goals behind it, the authors of the database are simply not allowed to make this set of recordings available to everyone. They received quite a few requests of artificial-intelligence (Big-Data) companies to share data with them, which clearly looked like attempts at appropriating data for commercial uses under cover of doing fundamental research. Companies may hide themselves behind individual researcher names in order to approach data owners. The database manager receives e-mail from Dr. XYZ to share the data for research; an internet search reveals that that person works for company ABC. Speech data, and annotated datasets in particular, can have high commercial value, and it does not sound right to allow companies to get hold of our data for free through permissive open-access licenses and make money using them without reciprocating for the corpus creators' generosity. In such cases, we need a framework that protects not just "our" data, but also us from being overtaken and made superfluous by large companies.

The argument concerning such data sets is compelling. To repeat a point already mentioned in the Introduction, research data need to be as open as possible, but also as closed as necessary. It is for each of us to ponder to what extent the data sets that we record fall into this category. It may turn out that this is by far not the majority case in linguistic research. Consultants' concerns about data being made public are not unwarranted: they reflect a growing awareness of widespread misuse of data. Vast amounts of data are being gathered by governments and companies, for debatable purposes. Data collection methods are sometimes illegal, and sometimes legal but unconstitutional (29). It is important to be aware of these developments, and to do what we can to deflect them. However, locking up our phonetic data sets does not look like a promising strategy to curb troubling trends in worldwide mass surveillance. Data-greedy companies and governments are not significantly impacted (it needs to be kept in mind that their data collection methods do not primarily consist in harvesting what is out there on the internet), whereas we suffer immediate consequences: we thereby deprive ourselves of resources for our research.

Crucially, issues of access to data need to be distinguished from the topic of data curation and archiving. Data curation and archiving can become standard in our field without depositors losing control of decisions about data access. (This point is taken up in §3.3 below.)

2.4 Data Preparation Makes Trouble

A further concern is that sharing of data requires an expense of time and resources which could place an individual scientist at risk: falling behind in research because one is working hard on data preparation for archiving. There is no denying that data preparation takes time and effort. But it does not have to be seen as an unpleasant additional requirement on top of other obligations. It can be viewed as an opportunity to convince reviewers and readers better, and to facilitate further progress (by the author and by others) by opening the data to other uses. Here is how this argument was stated at the 2017 Linguistic Society of America Institute Workshop on Data Management Plans:

> The rising tide of data management and sharing requirements from funding agencies, publishers, and institutions has created a new set of pressures for researchers who are already stretched for time and funds. While it can feel like yet another set of painful hurdles, in reality, the process of creating a Data Management Plan (DMP) can be a surprisingly useful exercise, especially when done early in a project's lifecycle. Good data management, practiced throughout one's career, can save time, money, and frustration, while ultimately helping increase the impacts of research.[v]

It may be good to recall here that there can be different levels of curation. A bare minimum (archiving the files securely, with indispensable metadata such as the Dublin Core set) already counts as a highly significant achievement, allowing others to take up the work later if they have the time and interest. Raw files accompanied by non-OCR'd scans of notes are infinitely more valuable than an absence of data (see, again, 2).

2.5 Individuals and Small Groups Cannot Change the Rules of the Game

A successful transition in science cannot be achieved by individuals. An expert who declines to review a manuscript or a dissertation on the grounds that there is no available data tends to be considered a troublemaker, putting a stain on their professional reputation, with no gain for the field. Some editors agree on principle that data are necessary for a fair and comprehensive review but consider that it would be unfair to the author of a specific manuscript to ask for access to audio files at the request of a reviewer: the argument is that lack of access to data currently constitutes such a widespread state of affairs that no author should be subjected to requirements which are not yet generalized. Journals have the power to be game changers, but editorial teams could worry that, should they decide to require data files accompanying new submissions, their journal would get fewer submissions: authors would favour venues that do not yet have this requirement. The additional workload is another deterrent. Clearly, the way forward is in collective decisions.

3 Recommendations

3.1 The Main Recommendation, and Its Justification

Our main recommendation is that journals publishing original research articles about phonetics decide that archiving of primary data is now a prerequisite for submissions. It would seem

advisable to deposit the data in a repository, and to link data to publications through references (using stable identifiers) rather than having the data tied to the journal, whose publisher may not be in a position to offer long-term archiving.

Importantly, mandatory data curation and archiving is distinguished from public access to data. Authors shall be encouraged to provide open data when possible. In the many cases where there are reasons to keep data closed, authors shall provide a public explanation why open access to data is not possible, and provide details on planned changes in access rights in future (see §3.3 below).

This recommendation essentially amounts to implementing the Peer Reviewers Openness (PRO) initiative (3) in the field of phonetics. A similar Open Science turn is being taken in various scientific fields: thus, the *Journal of Personality and Social Psychology* has taken this epoch-making move (30).

An alternative could be to build a parallel system to incentivize data archiving. Cogent proposals to this effect have been put forward by linguistic fieldworkers: assessing annotated corpora as research output (see in particular 31). But the practical implementation of this plan raises thorny issues. If data archiving were recognized as an academic achievement in itself, independently from research publications, how would the two be weighed when assessing researchers and academic institutions? Within a dual reward system, it is likely that most institutions would rate achievements in terms of publications higher than achievements in terms of data curation. Thus, the time invested into data curation and archiving would still yield lower returns in terms of career building, and so these tasks would not be efficiently encouraged.

Moreover, different subfields of linguistics have different traditions. The proposal to reward data curation and archiving as research output has been put forward by linguists who belong to the group of linguistic subfields now referred to as "diversity linguistics", which includes "descriptive linguistics (especially of previously understudied languages, often in a feldwork setting), language typology, and comparative linguistics" (32). Language documentation and conservation has long been recognized as an important goal by diversity linguists: collections of texts with interlinear glosses are one of the three pillars of linguistic fieldwork, along with dictionaries and grammars (33,34). Fieldworkers' sustained personal contacts with communities speaking endangered languages nurture a sense of the value and uniqueness of the data sets collected, as well as a sense of personal responsibility for passing on the data to future generations.

The situation of the phonetic sciences is somewhat different, with looser ties to language documentation. Giving additional credit to researchers for the archiving of phonetic data sets would feel like mixing apples with oranges. The heterogeneity could be partly covered up by using citation metrics as a uniform system to give credit for publications, data and tools. Using identifiers such as DOIs for archived data sets and for software as well as for publications is a way to give credit to all three within the same reward structure: if colleagues use data, software or publications, they provide a citation, and recognition ensues. But citation metrics are more of a problem than a solution: as *publish or perish* shifts to *impact or perish*, citation clubs and other scientific scourges follow (35). "All metrics of scientific evaluation are bound to be abused" (36). Clearly, establishing a more balanced reward structure for all research outputs requires a change in the academic incentive system altogether (once more, we are repeating a point made eloquently by 2). But if data curation is simply considered as an obligatory part of research output, then there is no need for separate reward structures, because curation and archiving are steps along the path to publication.

Another alternative would be to wait for funding agencies (or employers: both universities and companies) to take the initiative to make data curation and archiving an

obligation. It may be that appeals from within the community of scientists have no real power to curb data loss, and that data curation will only become common practice as it becomes an obligation enforced from above. (This would in a sense be similar to scientists' appeals to reverse worldwide environmental destruction (37), which, however cogent, remain a dead letter unless they are translated into political decisions.) A push in this direction has been under way for some time, on a country-by-country basis. The United States' National Science Foundation requires Data Management Plans ("are you making your data available, and how; if not, why?"), and European countries are gradually following suit. This is not necessarily an exciting prospect for our community, however. First, "if researchers do lip-service to an administrative requirement rather than strive for high quality to satisfy peer-review standards the outcome will not be transparency of the research process and safeguarding of unique data" (31). Said differently, if we do not really feel strongly about the usefulness of data sharing, and only archive data because we are requested to comply with regulations imposed on us, the important tasks of data curation are unlikely to be carried out with a sufficient amount of care and attention. "Archive stuffing" could become a sad equivalent (for research data) of "consent washing"[vi] in the field of ethics: eliciting consent according to regulations to protect oneself and one's employer from liability, without actually engaging with key ethical issues in their social and cultural context.

Moreover, differences between the frameworks in different countries and institutions create a confused landscape for our highly international and mobile community. Proposals produced from inside the community of researchers in the phonetic sciences appear as a much more promising way to go: planning the future of the field and advocating our vision as a community, rather than letting changes happen and having to manage the ensuing confusion and contradictions. We can debate among ourselves to choose reasonable strategies that make it possible to choose sharing over secrecy while minimizing potential problems: strategies that can be applied without inordinate expense of time and resources.

Concerning the implementation of this recommendation, two points are emphasized below: the need to rely on institutional archives (§3.2) and the need to set a date (however distant) for opening up the data.

3.2 Data Hosting in an Institutional Archive

The acronym FAIR summarizes four guiding principles for scientific data management and stewardship: data should be Findable, Accessible, Interoperable and Reusable (38). But it does not cover the important issue of maintaining access as time goes by. Making files available online through a website achieves the synchronic goal of allowing access to reviewers, readers and interested colleagues, but raises concerns about permanence. For the data to remain available in future, they should be hosted in an institutional archive that has provisions for long-term conservation. Open Science Framework (https://osf.io) and Zenodo (https://zenodo.org/) are examples of such archives. Open Science Framework (OSF) allows us not only to store data and materials, but also to cross-reference projects and link them to other platforms. This platform is tailored to scientists' use and appears well worth advertising. Zenodo offers flexibility for deposits: all file formats are accepted, and depositors are not requested to provide detailed metadata. Such flexibility will be welcomed by many phoneticians who have files in highly specific formats. A drawback of Zenodo's flexibility is that there is no guarantee against "data lock-in" (data in unusual formats becoming inaccessible for want of sufficient documentation about its format). On the other hand, projects such as Software Heritage (https://www.softwareheritage.org/) hold promise for preserving the necessary tools and information for motivated users to retrieve the data. Among archives specifically tailored for the

needs of linguists, many follow the recommendations of the Open Language Archive Community (OLAC: see http://www.language-archives.org/archives), so that the data sets deposited in any of these archives automatically appear in OLAC's lists of resources by language. Archives that participate in OLAC include the Linguistic Data Consortium (LDC), an available repository for corpora of audio recordings.

A further advantage of connecting to an institutional archive is that depositors can avail themselves of the archive's experience not only in terms of metadata management, but also in terms of consultants' consent and access rights (about which see also §3.3). Archives keep participants' consent forms on file – and those that do not do so as yet will gradually come to offer a management system for such 'meta-materials' – so that informed consent in written, audio or video form can be deposited alongside the primary data. Managing metadata (in a broad sense that includes participants' statements) requires know-how. This is no trivial task: heritage materials typically have no accompanying consent forms, and current practices and requirements vary from one institution to another. To address this important challenge, it is advisable to have professional archives as partners and allies, for reliability, stability and interoperability.

An alternative could be to have data hosted inside the same infrastructure as research publications. For instance, the online instructions for publishing in the journals published by the Linguistic Society of America (such as the flagship journal *Language*) are clear and simple, making it easy for authors to provide supplementary materials that support the research.[vii] In theory, such joint hosting could be a way to ensure close association of data to publications. But it places a high burden on the publishers. Offering adequate infrastructure for large data sets, technical support to authors, and provisions for long-term archiving could be seen as falling outside the scope of a publisher's core missions. Moreover, if data were hosted by the commercial publishers who currently retain a strong position in phonetics – such as Elsevier's *Journal of Phonetics* – valuable data sets would end up behind paywalls, raising further issues. Another alternative would be to set up a new archive specifically for phonetic/phonological data. Setting up an archive and making provisions for maintaining it for the long term is a huge amount of work, however. Having an archive of our own need not be seen as a prerequisite for archiving our data: it makes excellent sense to avail ourselves of existing institutional archives. The time and effort that we wish to devote to promoting Open Science are best placed in formulating clear guidelines for data curation, such that we can easily reuse one another's data.

3.3 Setting a Date for Opening up the Data

As recalled at the outset of §2.3, we phoneticians are accustomed to keeping our research data private, without a commitment to making them available to others at any point. The usual outcome is that the data are eventually lost. If, following the present proposal, the full data set is archived by the stage an article is submitted for publication, the new state of affairs will constitute a huge progress over the current situation. Access to the archived data is a distinct matter. A sample can be made available to reviewers, while the researchers reserve exclusive rights (keeping the data private) to make the most of the full data set in research before others are given a chance to re-use it. The duration during which data are kept private is a matter which we believe should remain for the depositors to work out, taking into account a variety of factors that include concerns of privacy (which differ widely from one type of materials to another) as well as the policies of employers and funding institutions. Adopters of Open Data principles (39) can choose to live out their commitment to immediate, unrestricted availability of data (with due attention to the language consultants' wishes), and others can opt for five or ten years of exclusive rights. A barrier of 25 or 50 years is likely to answer most concerns of privacy. A

duration of 100 years can be managed by archivists: thus, the French *Archives nationales* only allow public access to certain types of documents (from birth registers to technical specifications for weapons of mass destruction) after one hundred years.

In detail, practices are likely to continue to vary in time and space. Researchers' decisions depend on a host of factors which include the degree of emphasis placed on open-science criteria by thesis committees, tenure committees and other assessment bodies. The key point in our view is that the standard practice becomes for full data sets to be archived (with sufficient metadata that someone other than the author can understand and use the data set in future), with provisions for opening up at some later date. The date can be very distant, so long as enough curation has been done that someone else can make sense of the data later.

4 Concluding Note

The argument made here is no more than a reminder of core values in our field, as in science in general. Open Science is just science done right. It is in our interest to provide for, and support, "much-improved data curation" (40). Coherent open data policies in phonetics will place us in a strong position to address efficiently, as a field, the complex and sometimes contradictory injunctions which we receive from employers, funders and institutional partners. Clearly, a collective turn towards Open Science in phonetics would have massive benefits for science.

ACKNOWLEDGMENTS

Many thanks to Pat Keating, Fangzhi Jia, and the editorial team of the Journal of Speech Sciences for comments on draft versions. We also wish to thank the colleagues, too many to name, with whom we discussed the topic of data archiving over the years. The views expressed are the authors' responsibility.

REFERENCES

1. Demolin D. *Experimental methods in phonology.* TIPA Travaux interdisciplinaires sur la parole et le langage. 1992;28. Available from: http://journals.openedition.org/tipa/162

2. Berez-Kroeker AL, Gawne, Lauren, Kung, Susan Smythe, Kelly, Barbara F, Heston, Tyler, Holton, Gary, et al. *Reproducible research in linguistics: a position statement on data citation and attribution in our field.* Linguistics. 2018;56(1):1–18.

3. Morey RD, Chambers CD, Etchells PJ, Harris CR, Hoekstra R, Lakens D, et al. *The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review.* Royal Society Open Science [Internet]. 2016;3(1). Available from: http://rsos.royalsocietypublishing.org/content/3/1/150547.abstract

4. Nosek BA. *Center for Open Science: strategic plan.* OSF Preprints [Internet]. 2017. Available from: doi:10.31219/osf.io/x2w9h

5. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. *A manifesto for reproducible science.* Nature Human Behaviour. 2017;1:21.

6. Roettger TB. *Researcher degrees of freedom in phonetic research.* Laboratory Phonology: Journal of the Association for Laboratory Phonology. 2019;10(1). Available from: http://www.journal-labphon.org/articles/10.5334/labphon.147/

7. Lipton ZC, Steinhardt J. *Troubling trends in Machine Learning scholarship.* acmqueue - Association for Computing Machinery. 2019;17(1):1–33.

8. Roettger TB, Winter B, Baayen H. *Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility.* Journal of Phonetics. 2019;73:1–7.

9. Piwowar HA, Vision TJ. *Data reuse and the open data citation advantage.* PeerJ. 2013;1:e175.

10. McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, et al. *Point of view: How open science helps researchers succeed.* Elife. 2016;5:e16800.

11. Houtkoop BL, Chambers C, Macleod M, Bishop DV, Nichols TE, Wagenmakers E-J. *Data sharing in psychology: A survey on barriers and preconditions.* Advances in Methods and Practices in Psychological Science. 2018;1(1):70–85.

12. Borgman CL. *Big data, little data, no data: scholarship in the networked world*. MIT Press; 2015.

13. Simmons JP, Nelson LD, Simonsohn U. *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.* Psychological Science. 2011;22(11):1359–66.

14. Wagner P, Trouvain J, Zimmerer F. *In defense of stylistic diversity in speech research.* Journal of Phonetics. 2015;48:1–12.

15. Vasile A, Guillaume S, Aouini M, Michaud A. *Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger*. I2D - Information, données & documents. 2020; Available from: https://halshs.archives-ouvertes.fr/halshs-02870206

16. Nicenboim B, Roettger TB, Vasishth S. *Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German.* Journal of Phonetics. 2018;70:39–55.

17. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey EC, et al. *Many analysts, one dataset: Making transparent how variations in analytical choices affect results* [Internet]. PsyArXiv; 2017. Available from: https://osf.io/qkwst

18. Remijsen B, Ayoker OG. *Contrastive tonal alignment in falling contours in Shilluk*. Phonology. 2014;31(3):435–62.

19. Barnes J, Veilleux N, Brugos A, Shattuck-Hufnagel S. *The interaction of timing and scaling in a lexical tone system: an example from Shilluk.* In: Calhoun S, Escudero P, editors. Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences). Melbourne: Australasian Speech Science and Technology Association; 2019.

20. Tennant JP, Crane H, Crick T, Davila J, Enkhbayar A, Havemann J, et al. *Ten hot topics around scholarly publishing*. Publications. 2019;7(2):1–24.

21. Open Science Collaboration. *Estimating the reproducibility of psychological science.* Science. 2015;349(6251):aac4716.

22. Brugger P. *From haunted brain to haunted science: A cognitive neuroscience view of paranormal and pseudoscientific thought.* In: Houran J, Lange R, editors. Hauntings and poltergeists: Multidisciplinary perspectives. 2001. p. 195–213.

23. Bachelard G. *La Formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective*. Paris: Vrin; 1938.

24. Nickerson RS. *Confirmation bias: A ubiquitous phenomenon in many guises.* Review of General Psychology. 1998;2(2):175–220.

25. Fischhoff B. *Hindsight-foresight: The effect of outcome knowledge on judgment under uncertainty*. American Psychological Association; 1974. Available from: http://doi.apa.org/get-pe-doi.cfm?doi=10.1037/e459202004-001

26. Maurel L. *Quel statut pour les données de la recherche après la loi numérique ?* [Internet]. S.I.Lex. 2016. Available from: https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/

27. Niebuhr O, Michaud A. *Speech data acquisition: the underestimated challenge.* KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik. 2015;3:1–42.

28. Niebuhr O, Michalsky J. PASCAL and DPA: A *pilot study on using prosodic competence scores to predict communicative skills for team working and public speaking*. In: Proceedings of Interspeech 2019. Graz; 2019. p. 306–10.

29. Snowden E. *Permanent record*. Macmillan; 2019.

30. Leach CW. Editorial. *Journal of Personality and Social Psychology*. 2019; Available from: http://doi.apa.org/getdoi.cfm?doi=10.1037/pspi0000226

31. Thieberger N, Margetts A, Morey S, Musgrave S. *Assessing annotated corpora as research output.* Australian Journal of Linguistics. 2016;36(1):1–21.

32. Drude S. Reflections on diversity linguistics: Language inventories and atlases. In: McDonnell B, Berez-Kroeker AL, Holton G, editors. *Reflections on language documentation 20 years after Himmelmann 1998*. 2018. p. 122–31. (Language Documentation & Conservation Special Publication).

33. Boas F. *Tsimshian texts*. Washington: Government Printing Office; 1902. 244 p. (Bulletin of the Smithsonian Institution. Bureau of American Ethnology).

34. Woodbury T. *Defining documentary linguistics.* In: Austin P, editor. Language documentation and description. London: School of African and Oriental Studies; 2003. p. 35–51.

35. Baccini A, De Nicolao G, Petrovich E. *Citation gaming induced by bibliometric evaluation: A country-level comparative analysis.* PLoS One. 2019;14(9):e0221212.

36. Biagioli M. *Watch out for cheats in citation game.* Nature. 2016;535(7611):201.

37. Ripple WJ, Wolf C, Newsome TM, Galetti M, Alamgir M, Crist E, et al. *World scientists' warning to humanity: a second notice*. BioScience. 2017;67(12):1026–8.

38. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. *The FAIR guiding principles for scientific data management and stewardship*. Scientific Data. 2016 Mar 15;3:160018.

39. Lust BC, Blume M, Pareja-Lora A, Chiarcos C. *Development of Linguistic Linked Open Data resources for collaborative data-intensive research in the language sciences: An introduction.* In: Cambridge, MA: MIT Press; 2019.

40. Hanson B, Sugden A, Alberts B. *Making data maximally available*. Science. 2011;331(6018):649.

[i] "Heard melodies are sweet, but those unheard / Are sweeter; therefore, ye soft pipes, play on; / Not to the sensual ear, but, more endear'd / Pipe to the spirit ditties of no tone" (John Keats, "Ode on a Grecian urn")

[ii] http://research-ethics.org/topics/data-management/?print

[iii] UC San Diego Resources for Research Ethics Education, http://research-ethics.org/topics/data-management/#regulations-and-guidelines (consulted April 5th, 2020).

[iv] Some will be surprised to learn that some countries make it a legal obligation to share research data. In French law, a literal reading of the "Digital Republic" law *(loi n° 2016-1321 du 7 octobre 2016)* is that data produced by public institutions, including research institutions, need to be made open (26).

[v] https://sites.google.com/a/hawaii.edu/data-citation/lsa-2018-workshop Similar statements are being steadily issued by other institutions. See, for instance, Unicamp's Institutional Policy of Open Access: http://repositorio.unicamp.br/static/sobre.jsp?locale=en

[vi] https://scinfolex.com/2020/05/01/stopcovid-la-subordination-sociale-et-les-limites-au-consent-washing/

[vii] https://www.linguisticsociety.org/sites/default/files/Supplemental%20Materials%20Guide%20for%20LSA%20Pubs_0.pdf, consulted Feb. 1st, 2020.