Looking for hidden speech archives in Italian institutions

Vincenzo Galatà
Institute of Cognitive Sciences and Technologies, National Research Council, Italy and University of Siena, Italy
vincenzo.galata@pd.istc.cnr.it

Silvia Calamai
University of Siena, Italy
silvia.calamai@unisi.it

Abstract

The paper presents the aims and the main results of an on-line survey concerning speech archives collected in the fields of Social Sciences and Humanities among Italian scholars. A huge amount of speech archives is preserved among researchers: most of the resources are not accessible and legal issues are generally not addressed in detail. The great majority of the respondents would agree to storing their archives in national repositories, if any.

1 Introduction

Very few surveys describe the amount and the size of speech archives in Italy. To our knowledge, only Barrera et al. (1993) and Benedetti (2002) map the existing audio archives. The first survey was made under the aegis of the Ministry of Cultural Heritage and listed only the public archives (Barrera et al. 1993). Benedetti (2002) listed also private archives, especially in the field of music. Sergio (2016) presented the photo and audiovisual archives that were digitised (or were in the process of being digitised) by public and private institutions in Italy. Other surveys were limited to a single area, such as AAVV (1999), devoted to Piedmont region, and Andreini & Clemente (2007) and Cappelli & Rioda (2009) which restricted the survey to the Tuscany. Partial inventories can be found scattered around the internet, especially within the context of the "Istituti Italiani per la Resistenza". It has to be noted that the great majority of the inventories focused on music and oral history archives and completely neglected the huge amount of material collected by linguists during their fieldwork. At the European level, an overview on the Oral History collections was made accessible and maintained by CLARIN ERIC[1]. At present, the overview contains about 260 collections scattered in 17 European countries (with great disparities between EU countries in terms of coverage and details). As for Italy, 86 collections are listed (data were collected in 2016 by the second author together with the Italian Association for Oral History).

The present paper aims at providing an updated map of Italian speech archives generated by field researches within and outside the academia, especially in the areas of linguistics and oral history, but also in other areas that we included while the survey was already running. Most of the archives we discovered are inaccessible and can be labelled as audio 'legacy data': that is, data stored in obsolete audio media by individual researchers outside of archival sites such as libraries or data centres. For this purpose, we set up an online survey in order to:

i) draft a survey of institutional archives, that is a survey of the existing speech archives deposited in (and by) institutions and associations;
ii) draft a survey of the existing speech archives owned by single researchers;
iii) provide an extensive analysis of the existing practices of collection, preservation and reuse in order to give a detailed description of the state of conservation and accessibility, the access policies, costs and sustainability.

The survey also made it possible to verify how the knowledge of the CLARIN infrastructure is widespread among Italian research communities. A bottom-up approach, involving the main Italian scientific associations, allowed us to reach as many researchers as possible and to bring a hidden, inaccessible, endangered treasure to light.

The paper is conceived as follows: §2 presents the structure and the content of the questionnaire prepared to run the survey; §3 reports on the sample that answered to the survey together with the main

[1] http://oralhistory.eu/collections/clarin-eric.

Vincenzo Galatà and Silvia Calamai 2019. Looking for hidden speech archives in Italian institutions. *Selected papers from the CLARIN Annual Conference 2018.* Linköping Electronic Conference Proceedings 159: 46–55.

results; §4 addresses some concluding remarks and underlines the urgent need to find an Italian repository to host these materials.

2 The questionnaire

The survey was jointly designed by both authors and was administered in Italian through an online questionnaire (implemented via Google Forms). The following Italian scholarly organisations were involved in the dissemination of the survey by means of their respective mailing list: *CLARIN-It*, *Associazione Italiana di Scienze della Voce* (AISV), *Associazione Italiana di Storia Orale* (AISO), *Società di Linguistica Italiana* (SLI), *Associazione di Storia della Lingua Italiana* (ASLI), *Associazione Italiana di Linguistica Applicata* (AITLA), *Società Italiana per la Museografia e i Beni Demoetnoantropologici* (SIMBDEA), *Associazione Italiana di Sociologia* (AIS), *Società Italiana di Antropologia Culturale* (SIAC), *Società Italiana di Antropologia Applicata* (SIAA), *Associazione Nazionale Professionale Italiana di Antropologia* (ANPIA). Other formal and informal networks were targeted (e.g. *Analisi dell'Interazione e della Mediazione* network, AIM) and also individual researchers – both in Italy and abroad – were personally contacted by email. We can presumably assume that several hundred scholars were reached by the survey.

The questions included in the survey were mostly yes-no and multiple response type (for three questions, for which it was impossible to predict or suggest any option, open answers were allowed). The questions were as generic and as inclusive as possible in order to be answered by all of the respondents and thus avoiding to focus on very specific scientific domains. In most of the cases, besides every yes-no or multiple response question, an "*Other, please specify*" field was provided in order to account for responses not foreseen by the authors. The choice of including such an open-ended response option had the disadvantage of increasing the amount of post-processing needed at the time of results reporting for each question (the responses resulting in highly scattered distributions), but at the same time this allowed the authors to account for multiple domains and issues not previously considered.

The survey was administered in Italian and was structured according to four distinct sections:

1) the first section was mainly informative and preceded the questionnaire itself by providing a brief presentation of the aims and the scope of the survey, as well as general information on the treatment of the recorded responses to the questionnaire;
2) the second section (displayed in the Appendix) contained the actual questionnaire consisting of 19 questions. The first question (Q.0) gave the participants the possibility to opt-out from the survey (thus registering their participation) or eventually to contribute to the survey, without necessarily completing the survey, by jumping to the third section of the survey (see point 3). The core questionnaire, consisting of 17 questions (Q.1 to Q.17), was devised in order to obtain a rough description and quantification of audio-visual resources (also with respect to accessibility and legal issues). One last question (Q.18) asked the respondents if they were aware of the existence of the CLARIN European infrastructure. A translated version of this section is provided in Appendix at the end of the paper;
3) the third section allowed all the respondents to contribute to the survey dissemination by suggesting the authors further potential contacts they considered worth to be contacted;
4) the last section of the questionnaire asked the respondents for some personal information (contact, academic position and affiliation).

For the aim of the current paper, in the next paragraph we report the results from selected questions of the survey, leaving the rest and more elaborate analyses to an extended version on the same topic. The questions on which we focus here are intended to:

1) uncover the scientific domains with the highest amount of hidden spoken resources;
2) identify what sort of resources we are coping with;
3) understand if digitised data (such as transcriptions, annotations etc.) are eventually available for these resources and in what format they are stored;
4) establish if the mentioned resources are accessible and who is in charge of their maintenance;
5) take stock of the ethical issues related to the creation of the resources under scrutiny;
6) assay how much the knowledge of the CLARIN European infrastructure is widespread in the different scientific domains.

3 Main results

The results we report on in this section refer to the responses gathered from the survey at the time of writing[2] with reference to selected questions as mentioned in the previous paragraph. So far, 151 respondents took part in the survey: 131 participants completed the survey, 17 opted-out and 3 only suggested other contacts. If we consider the affiliation of the respondents specified in the last section of the questionnaire, we reached 86 individuals declaring an affiliation in Italy and other 8 with affiliation either in Switzerland, Spain, UK, Norway, Belgium or Ireland (36 did not declare any affiliation).

Since for most of the questions the participants were allowed to select multiple responses and eventually to specify further responses on an extra field, if not otherwise stated we will always refer to the number and percentage of cases mentioned by our respondents.

3.1 Spoken resources and their scientific domains

One of the first questions of the survey (Q.1) asked the respondents to mention all of the scientific domains to which their resources belong to. Besides some possible options provided by the authors, the respondents had the possibility to report other domains in an optional field. The responses to this question have been grouped to form a sort of top list of domains as in Figure 1. The most mentioned domains in decreasing order are: *Oral History* (n = 53), *Phonetics & Phonology* (n = 35), *Dialectology* (n = 33), *Anthropology* (n = 31), *Ethnomusicology* (n = 8), *Language Acquisition* (n = 7), *Sociolinguistics* (n = 6), *Applied Linguistics* (n = 6), *Sociology* (n = 5), *Conversation Analysis* (n = 4), *Speech Technology* (n = 2).[3]

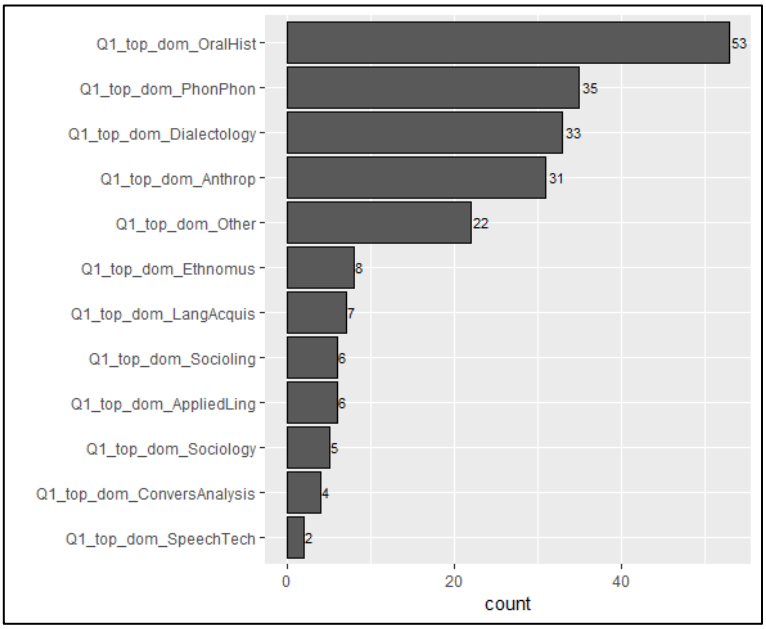

Figure 1. Scientific domains under which the respondents placed their resources.

After grouping the responses to this question into macro-areas[4], our initial intuition (e.g. that the huge amount of material collected by linguists during their fieldwork is neglected) stands out (see Figure 2). The majority of the participants we were able to reach indicated *Linguistics* (40.7%), *Oral History*

[2] The survey (available at https://goo.gl/8uHYK1) started on February 20th, 2018. Despite our initial intentions, the survey is still open and will be kept open until end 2019. This will allow the authors to continue the survey by reaching more respondents and eventually to include areas we might have neglected so far.

[3] The other *Linguistic* subfields mentioned in sparse order (e.g. prosody, syntax) and other hapax domains have been categorized as *Other* (n = 22).

[4] Due to the possibility the respondents had to fill in the "*Other, please specify*" option when indicating the scientific domains under which they considered their resources, the results on the disciplines were unavoidably scattered. To this end, following the *Linguistics* subfields grouping in the OLAC project (http://www.language-archives.org/REC/field.html), we recoded the responses to further reduce the sparseness of the data.

(30.8%) and *Anthropology* (18.0%) as core domains for their resources, with a minor portion of them indicating *Ethnomusicology* (4.7%), *Sociology* (2.9%), *Speech Technology* (1.2%), *Other* (1.7%).

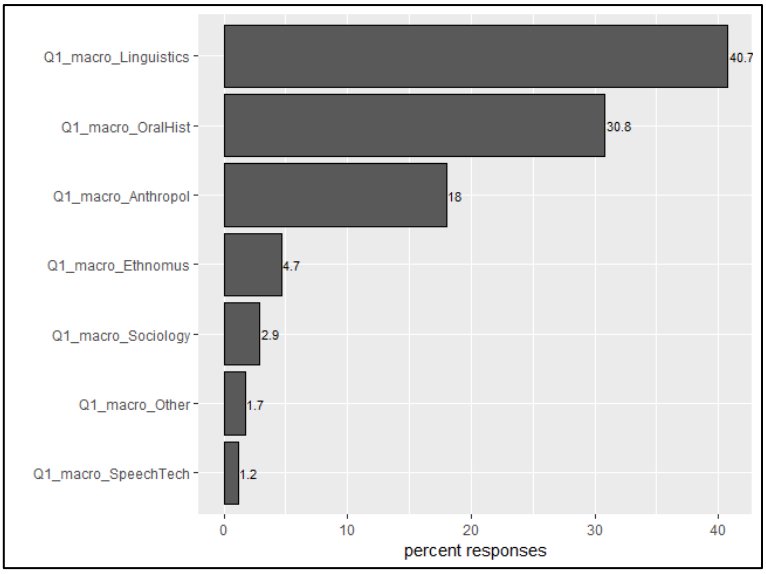

Figure 2. Macro-areas under which the respondents placed their resources.

3.2 Type of resources involved

When collecting speech in the different domains, the spoken productions can be recorded as uni-modal signals (e.g. Audio only) or as bi-modal signals (e.g. Audiovideo). This consideration led the authors to include this distinction in the survey (Q.2, see Figure 3). As few as 13% of the respondents selected Audiovideo only, while 40.5% of them declared having both Audio and Audiovideo resources. Those who declared having Audio only resources represent 46.6% of the cases.

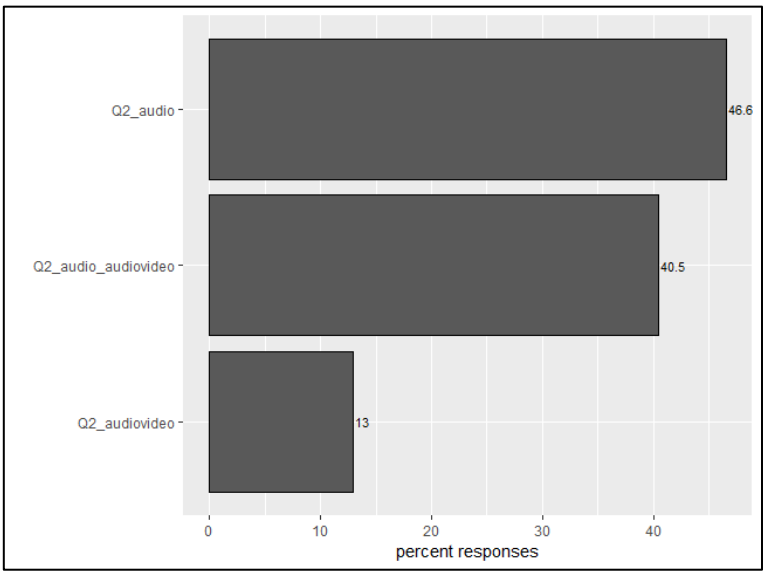

Figure 3. Type of resources owned by the respondents.

In the survey we further asked the respondents to indicate of what type of resources they were in possession of (Q.7), meaning here if these were of digital or analogue nature. As much as 70.5% of the resources were mentioned to be of digital nature (e.g. *.wav, *.Mp3, *.eaf, *.TextGrid, *.txt etc.), 26.7% of analogue nature (tapes, forms etc.), one respondent selected DAT (Digital Audio Tape) and other four did not answer the question. Additionally, for those who mentioned *Digital* in their answer, we asked to clarify the nature of those resources (Q.8). The options (and percent of responses in brackets) were:

a) born as original digital resources (54%);

b) the product of digitised analogue resources of which the respondents still own the originals (20%);
c) the product of digitised analogue resources of which they do no more own the originals (9%);
d) digitised "copies" of files owned by others (12%);
e) no answer (5%).

As our survey reveals, more than half the resources (54%) consists of original digital resources. It is undoubtedly a sign of the time: in the past, interviews "tended to be recorded on professional quality and somewhat bulky open-reel tape recorders" or, more recently, on "prosumer grade audio cassette recorders" (Cieri 2011: 31). Such venerable devices are now replaced by modern digital recording equipment which is nowadays within the reach of everyone.

Going more into detail, with Q.9 we also asked the respondents to tell us on what type of media the resources were stored: we gave them a list of predefined options with the possibility to select *everything that applies* and an additional "*Other*" option to specify other media types not listed. The responses are obviously scattered, but what is somehow surprising - if one looks at Figure 4 - is that only 28 of those who took part in the survey have their resources safely stored (*Server with back-up*).

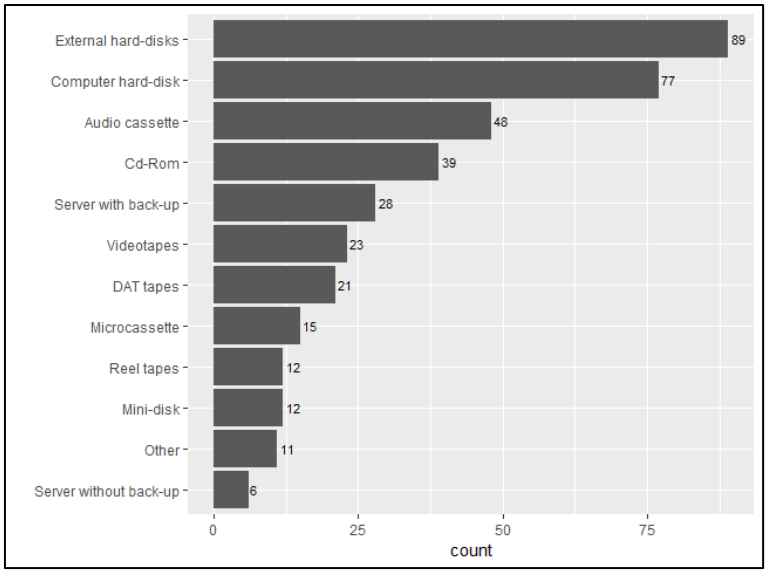

Figure 4. Media types on which the resources are stored.

The asked distinctions above are anything but trivial as they have direct consequences both on the quality and on the size (and eventually format) of the resources. In this respect, when it comes to quantifying the owned spoken resources (mainly in terms of hours) only half respondents (57%, 75 out of 131) are able to tell how much data they have in their archives (Q.6). Some of the respondents quantified their resources either in terms of number of interviews or number of files or number of tapes etc.; others were more precise and indicated an estimate in number of hours. Taking into account the responses gathered so far from the 75 respondents who were able to quantify their resources, and if we dare to do a very brutal conservative estimate of the amount of audio resources of those who were able to quantify it in terms of hours, the amount is impressing: more than 12 thousand hours of recorded material emerge from our survey, with more than 10 thousand hours mentioned (either exclusively or additionally) in the *Linguistics* domain.

The data above just provide a rough estimate giving us an idea of how much data is somehow remaining hidden to us. In addition to what was shown so far, we also asked to specify the language of their resources (Q.5). It might be of interest to know that 80.9% of the respondents answered having resources in *Italian language*, 42.6% listed *dialect varieties of the Italian peninsula* and 23.7% declared resources in *other languages*.

3.3 Type and format of additional data available for the targeted resources

Another question (Q.3) asked whether additional textual data in digital format (e.g. transcriptions, annotations etc.) are available besides the speech resources and, if there are, what type of format these data have (Q.4).

The great majority of the respondents (80.9%) declared additional data in digital format. As can be seen from Figure 5, considering all the mentioned types of files, the most common ones listed are *.doc (26.9%)[5], *.txt (24%), PRAAT's *.TextGrid (15.6%), ELAN's *.eaf files (7.2%), *.pdf (4.8%), Transcriber's *.tag files (1.2%) and of other sparse formats (4.8%).

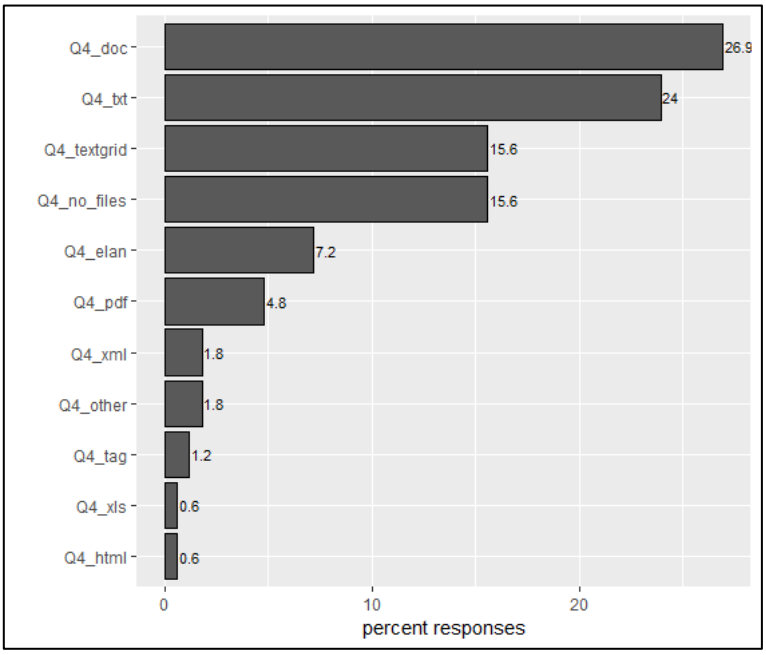

Figure 5. Declared additional textual data in digital format.

As far as the format of additional data available besides the speech resources is concerned, we categorized the above responses into two distinct groups: binary and non-binary files. This categorization was considered important also in order to verify if the information stored in those files is easily accessible and thus void of any restriction. Binary files are commonly application specific (e.g. proprietary) files. Due to the obsolescence of many applications, the use of binary files (as opposed to non-binary files, which allow unrestricted access and interoperability) has serious side-effects related to accessibility issues on the long term. Among the file formats, the respondents listed both *binary* (34%, including *.doc, *.pdf etc.) and *non-binary* formats (51%, including *.txt, PRAAT's *.TextGrid, ELAN's *.eaf, Transcriber's *.tag etc.); for 26 respondents (15%) no additional files are available.

3.4 Accessibility and maintenance issues

Accessibility of the data was addressed with Q.10 and it is not surprising to discover that almost half of the resources listed in our survey (49.6%) is barely accessible. Only 9.2% of the resources is accessible and available, 4.6% is partially accessible, 35.1% is available upon request, 1.5% is available upon request and only for selected parts. Moreover, we also wanted to know how one might have access to the resources they declared to be accessible and so we asked if specific access policies were available (Q.11). Only 9.2% of these resources is freely accessible online (with no authentication); 7.6% is accessible online via authentication; 29% is accessible onsite (i.e. where the resources are physically stored).

However, the two questions mentioned above open up to another important question: who is in charge of their long-term maintenance and preservation? For this reason we asked Q.12. Not surprisingly, the answer receiving the highest number of responses was *nobody* (43%), followed by *reference Institutes*[6] (17%), *reference Universities* (16%), the *owners/individuals* themselves (15%). Most surprisingly is the very small number of individuals (n = 5, i.e. 3.5%) who mentioned an *external repository* (NA's = 5.6%).

[5] For sake of economy, we grouped under the *.doc extension a series of other extensions such as *.docx, *.odt and *.rtf as well as all the generic responses (such as "word") referred to the popular word processor.
[6] Under this label we grouped institutions such as associations, foundations, libraries and their archives.

We would also like to stress the fact that the necessity of a national repository is of the highest urgency if we consider that most of those owning speech resources in our survey (about 47%) fall within the category which we defined as *casual workers* (e.g. workers without a permanent position or a permanent affiliation to an institution). Only 37% of the remaining respondents declared a *permanent* position and affiliation (for example to universities or other public institutions), while 9.2% did not provide any information (the remaining 7.6%, which we were not able to ascribe to any of the two categories, has been categorized as a generic "*other*" category).

3.5 Ethics and legal issues concerning oral resources

One further information emerging from our survey (Q.14) relates to ethics and legal issues, which are addressed by the respondents only in 46% of the cases.

Even if undoubtedly most part of the resources have been collected at a time when privacy and data protection issues were not addressed as nowadays, the effects on the accessibility and reusability of such resources are unavoidable. Not all the researchers are probably aware of new elements introduced by the recent General Data Protection Regulation (GPDR, EU n.2016/679), although certain scientific associations are providing information to their members, in order to support them in such a challenging issue (e.g. Italian Association of Oral History, http://aisoitalia.org/buone-pratiche-di-storia-orale-alcune-importanti-novita/). At the same time, the authority responsible for privacy in Italy has been organising several information meetings with Italian universities and public research bodies in order to raise awareness among the different research communities and university administrative staff on the changes introduced by the GDPR and their impact on research and dissemination activities.[7]

3.6 The CLARIN European infrastructure in our survey's scientific community

An unexpected result emerging from our survey at Q.18 is that only 31% of the respondents declared to have knowledge of the CLARIN infrastructure. This low percentage, however, should not discourage and diminish the activities carried out so far within the CLARIN infrastructure, on the contrary. There is indeed a large pool of resources owners (e.g. 64%, more than half of our respondents) who would agree to storing their archives and their speech resources in national repositories (Q.16). This manifestation of interest should give CLARIN's mission more strength and actuality.

4 Conclusion

In the past, researchers usually considered their speech data valuable only for the immediate purposes of their research. Nowadays, we are facing a change in attitude, since it is clear that legacy data document previous states of languages and linguistic changes from different points of view, and allow to work on historical questions about languages. Moreover, speech archives perfectly fit into the international debate concerning the use and reuse of past research data. Several scholars pointed out many important advantages of re-analysis: from sustainability to the maximization of the results. At the beginning of a novel research project, the re-analysis of past archives can be invaluable in providing a first orientation on the topics to be investigated, and therefore making the pilot stage of the research both more effective and swifter. By making previous research data available to re-analysis by others, it is possible to multiply the research outcomes through the publications of further interested scholars.

Nevertheless, the outcome of our survey shows a rather delicate picture: rather limited accessibility of the resources, ethical and legal issues only partially addressed, scant knowledge of the CLARIN infrastructure. In order to start filling the gap, the topic of the annual conference of the Italian Speech Sciences Association (AISV, *Associazione Italiana Scienze della Voce*) held in February 2019, was devoted to speech archives. The conference also saw the participation of the Executive Director of CLARIN who gave a keynote lecture exactly on *Spoken Word Archives as Societal and Cultural Data* (https://www.aisv.it/aisv2019/en/program).

[7] See for example https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/8318508 and https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/7977380 [date accessed: 26.01.2019].

Acknowledgements

The authors thank the anonymous reviewers for valuable comments on an earlier version of this paper, the organisations for their collaboration and all respondents for their participation in the survey.

References

Andreini A., Clemente P. (eds) 2007. *I custodi delle voci. Archivi orali in Toscana: primo censimento*, Firenze: Regione Toscana.

Barrera G. et al. 1993. *Fonti orali. Censimento degli istituti di conservazione*, Min. Beni Culturali e Ambientali.

Benedetti A. 2002. *Gli archivi sonori: fonoteche, nastroteche e biblioteche musicali in Italia*, Genova.

AA.VV 1999. *Archivi sonori. Atti dei seminari di Vercelli (22 gennaio 1993), Bologna (22-23 settembre 1994), Milano (7 marzo 1995)*, Roma, Min. Beni e le Attività Culturali-Ufficio centrale per i Beni archivistici, 1999.

Cappelli F., Rioda A. 2009. Archivi sonori in Toscana: un'indagine, Musica/Tecnologia, 3: 9-69.

Cieri C. 2011. Making a field recording. In *Sociophonetics*. *A student's guide*, M. Di Paolo, M. Yaeger-Dror (eds.), 24-35. Abingdon: Routledge.

Sergio G. (ed) 2016. *Atlante degli archivi fotografici e audiovisivi italiani digitalizzati*, Venezia: Fond. di Venezia-Marsilio.

Appendix

Questionnaire used in the survey (English translation)

Questionnaire - Section 2
All answers are optional. However, please try to answer as completely as possible, trying not to leave any questions.

0. Do you own any oral data / multimedia resources?
   Select only one.
   ○Yes
   ○ No (stop filling out the form and scroll down to the last question to terminate the survey)
   ○ No, but I know someone who owns some (scroll down and go to section 3)

1. Under which disciplinary area does your collection of oral data / multimedia resources fall?
   Select all that applies.
   □ Phonetics/phonology
   □ Dialectology
   □ Oral history
   □ Anthropology
   □ Sociology
   □ Psychology
   □ Applied linguistics
   □ Sociolinguistics
   □ Ethnomusicology
   □ Other, please specify: _____

2. Of what kind of resources is it about?
   Select all that applies.
   □ Audio
   □ Audiovideo

3. For those resources, do you own also textual files in digital format such as for example transcriptions, annotations etc.?
   Select only one.
   ○ Yes
   ○ No

4. If you answered “Yes” to the previous question, what format do these textual files have?
   Select all that applies.
   □ *.TextGrid (PRAAT)
   □ *.eaf (ELAN)
   □ *.tag (TranscriberAG)
   □ *.txt
   □ *.doc
   □ *.pdf
   □ Other, please specify: _____

5. In what language are these resources?
   Select all that applies.
   □ Italian
   □ Dialect varieties of the Italian peninsula (specify which in “Other, please specify “)
   □ Other languages (specify which in “Other, please specify”)
   □ Other, please specify: _____

6. Are you able to provide one or more estimates on the amount of multimedia resources you own?
   Please, if your answer is “Yes”, use the “Other” field and summarily describe the amount of such data as best you can. In reporting any numerical values, please also indicate the units of measurement to which you refer, for example in terms of gigabytes, hours, minutes, etc.
   Select all that applies.
   □ I am not able to quantify
   □ Yes
   □ Other, please specify: _____

7. The resources are in the following format:
   Select all that applies.
   □ Digital (*.wav, *.Mp3, *.eaf, *.TextGrid, *.txt etc.)
   □ Analogue
   □ DAT
   □ Other, please specify: _____

8. If you selected also “digital” in the previous question, can you please specify whether these resources are:
   Select all that applies.
   □ Born as original digital resources
   □ The result of digitised analogue resources of which I still own the originals
   □ The result of digitised analogue resources of which I no more own the originals
   □ Digital “copies” of resources owned by others
   □ Other, please specify: _____

9. Can you tell us on what type of media they are stored?
   Select all that applies.
   □ Audiotapes Microcassettes
   □ DAT tapes
   □ Video tapes
   □ Reel tapes
   □ Mini-disk
   □ Cd-Rom
   □ External hard-disk
   □ Computer hard-disk
   □ Server with back-up
   □ Server without back-up
   □ Other, please specify: _____

10. Are the data and the multimedia resources freely accessible?
Select only one.
□ Yes
□ Yes, but only upon request
□ No
□ Yes, but only partially
□ Other, please specify: _____

11. Are there specific access rules?
Select only one.
□ None (no access rules are available)
□ Online (with authentication)
□ Online (free access without authentication)
□ On-site
□ Other, please specify: _____

12. Who is in charge of the long-term maintenance and preservation of these resources?
Select all that applies.
□ Nobody
□ Reference University
□ External repository (specify which one in "Other")
□ Reference Institute
□ Owner
□ Other, please specify: _____

13. Where and how are your multimedia resources stored?
Please, provide as much detail as possible.
_____

14. Are the ethical and legal aspects of the data collection regulated (e.g., intellectual property, potential data reusability)?
Select only one.
○ Yes
○ No

15. If "Yes", how? Would you be willing to give us an example of a consent form normally used in your laboratory or in your research group (we will not divulge any form you will provide us)?
You can copy and paste the text of the form into the field below.
_____

16. Would you be interested in depositing also somewhere else your data?
Select only one.
○ Yes
○ No

17. If you answered "yes" to the previous question, what would you base your choice upon in deciding to deposit your resources on a potential repository?
For example, the presence of a graphical user interface, the presence of a dedicated structure offering support if needed, a free deposit service, the possibility to index the resources, etc.
_____

18. Do you know the CLARIN European infrastructure?
Select only one.
○ Yes
○ No