Looking for hidden speech archives in Italian institutions

Vincenzo Galatà[1,2] | Silvia Calamai[1]

[1]University of Siena, Italy | [2]Institute of Cognitive Sciences and Technologies, National Research Council, Italy

vincenzo.galata@pd.istc.cnr.it | silvia.calamai@unisi.it

with support from AISV Associazione Italiana Scienze della Voce

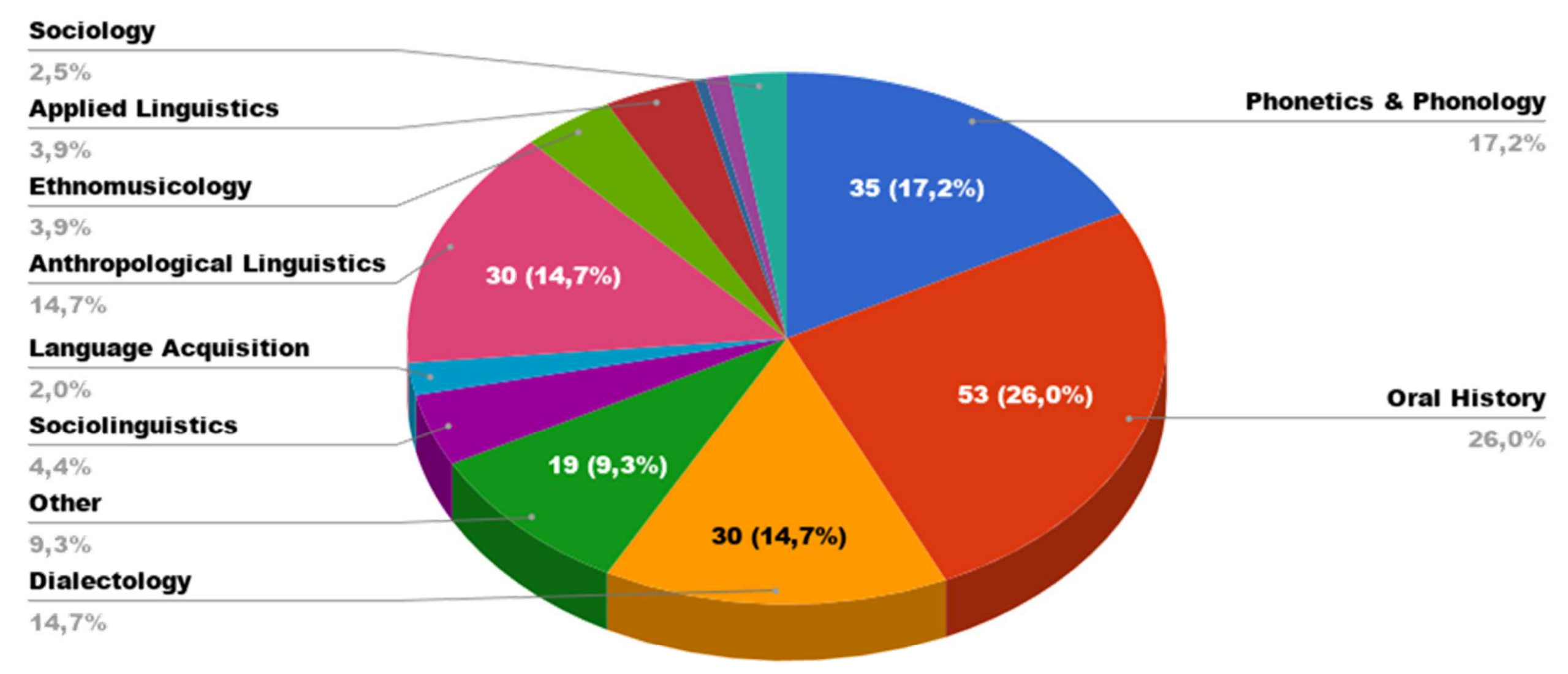
Abstract

By means of a **survey**, our aim is to **provide** 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 (most of the archives are unavailable and can be labelled as audio 'legacy data', e.g. data stored in obsolete audio media by individual researchers outside of archival sites).

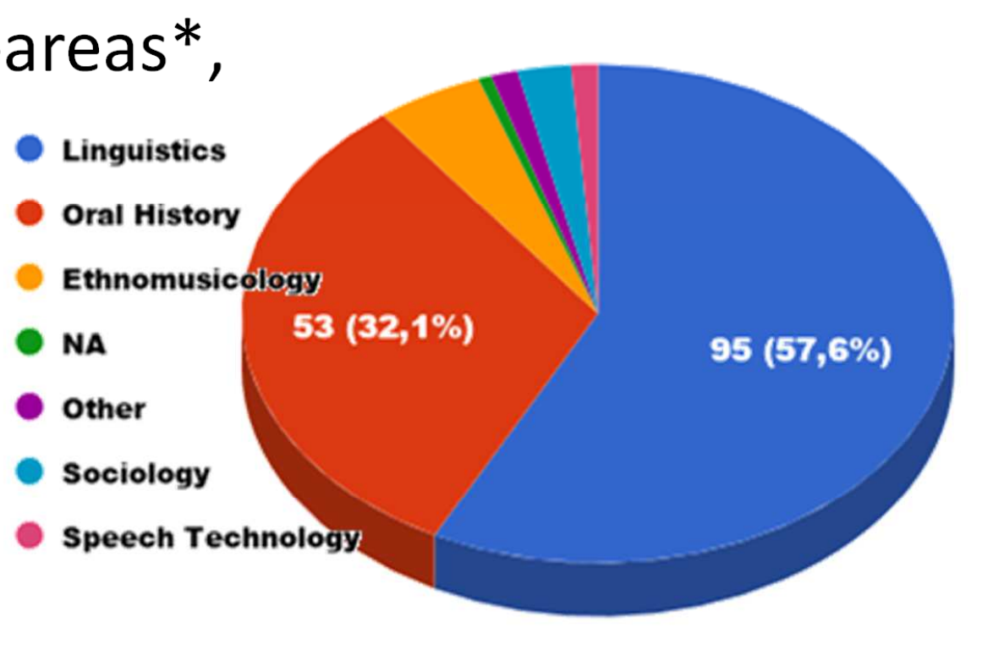
A **bottom-up approach**, involving the main Italian scientific associations (AISV, AISO, SLI, ASLI, AITLA, SIMBDEA), other formal and informal networks and personal contacts, allowed us to reach as many researchers as possible and to bring a hidden, inaccessible, endangered treasure to light.

A. Spoken resources & their scientific domains

The most mentioned scientific domains to which the resources belong to.



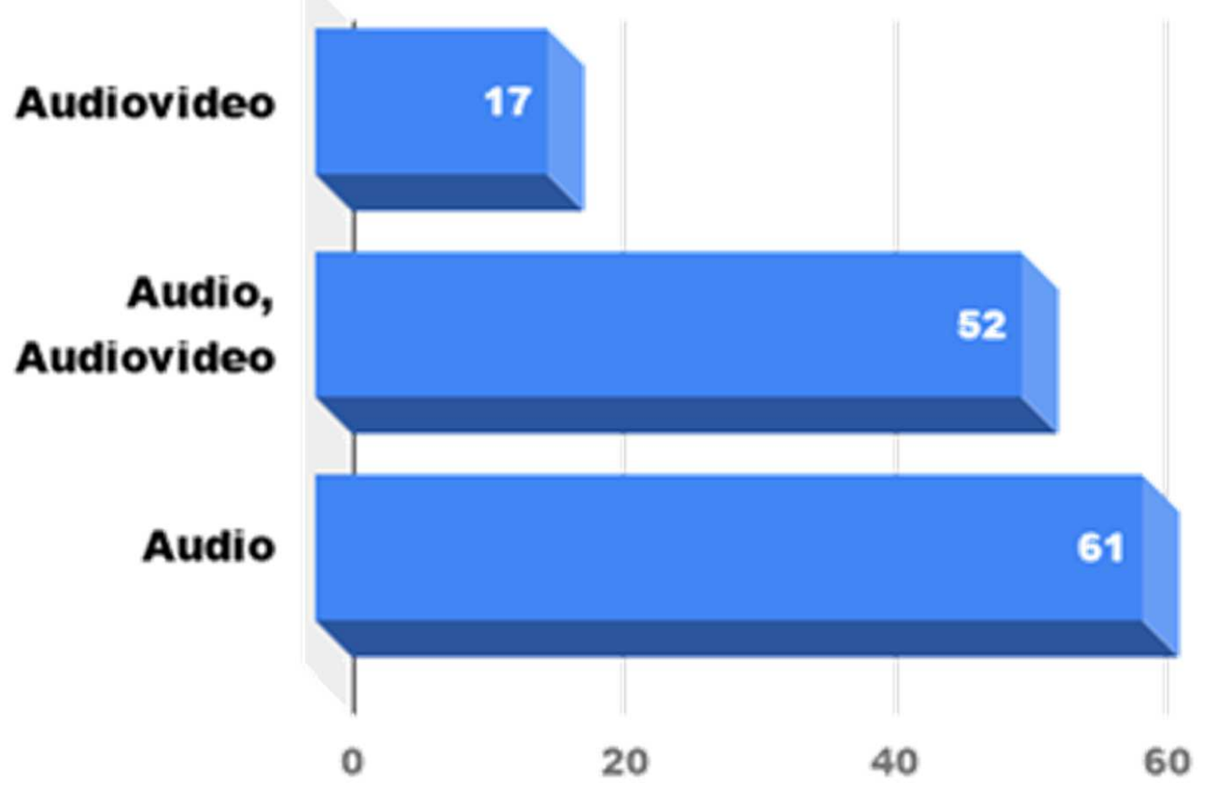
After grouping the same responses into macro-areas*, our initial intuition (e.g. that the **huge amount of data collected by linguists** during their fieldwork is neglected) stands out.



*Following the *Linguistics* subfields grouping in the OLAC project (http://www.lan-guage-archives.org/REC/field.html) we recoded the responses to reduce the sparseness of the data.

B. Type of resources involved

Spoken productions in the different domains can be recorded as a uni-modal signal (e.g. Audio only) or as a bi-modal signal (e.g. Audiovideo).

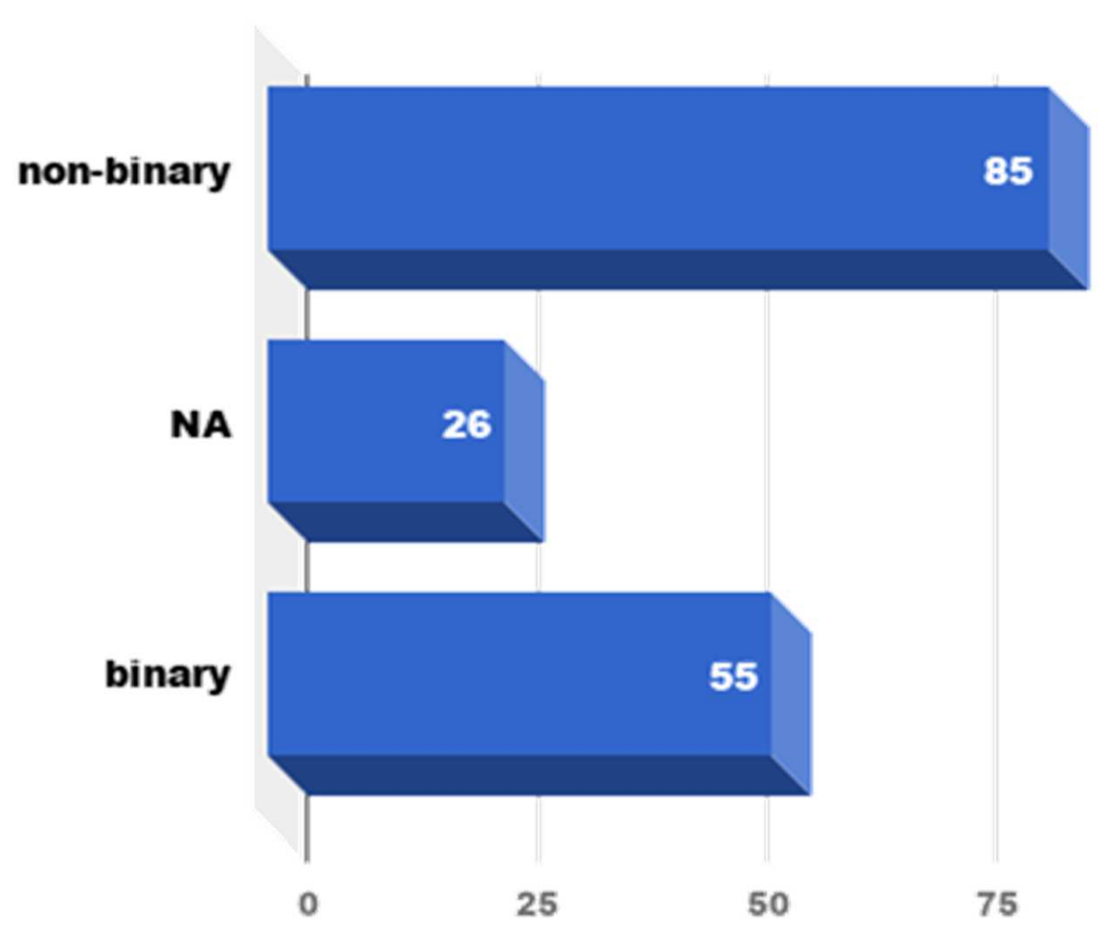


The respondents chose in **13.1%** of the cases **Audio-video only**, in **40% both Audio and Audiovideo**, and the remaining **46.9% Audio only**.

The asked distinction is anything but trivial as it has direct consequences both on the quality and on the size (and eventually format) of the resources.

We further asked the respondents to indicate of what type of resources they were in possession of: **70.7%** of the resources were mentioned to be of **digital nature** (e.g. *.wav, *.Mp3, *.eaf, *.TextGrid, *.txt etc.); **26.4%** of **analogue nature** (tapes, compact cassettes, etc.).

C. Type & format of additional data available



Responses to this Q categorized into **binary** and **non-binary** files in order to verify if the information stored is easily accessible and unrestricted.

The most common listed files are:

- **Binary**: *.doc (23.1%), *.pdf (4.7%)
- **Non-binary**: *.txt (23.1%), PRAAT's *.TextGrid (15.4%) and ELAN's *.eaf files (7.1%).
- 26 respondents did not specify any format despite declaring the presence of additional files for the speech resources.

Due to the obsolescence of many applications, the use of binary files (e.g. application specific and proprietary 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.

The survey (still available at https://goo.gl/8uHYK1)

Mostly **yes-no** and **multiple response** type questions (Qs) **as generic** and **as inclusive as possible** in order to be answered by all of the respondents ("*Other, please specify*" field provided in order to account for unforeseen responses).

The survey was structured according to **4 distinct sections**:

- **Section 1** - informative - brief presentation of aims and scope of the survey, as well as general information on the treatment of the collected responses;
- **Section 2** - the actual survey (**19 Qs**) with the possibility for the participants to opt-out by jumping to the 3rd section. The last question asked the respondents if they were aware of the existence of the CLARIN EU infrastructure;
- **Section 3** - respondents contribute to the survey dissemination by suggesting further potential contacts;
- **Section 4** - respondents' personal information (contact, academic position and affiliation).

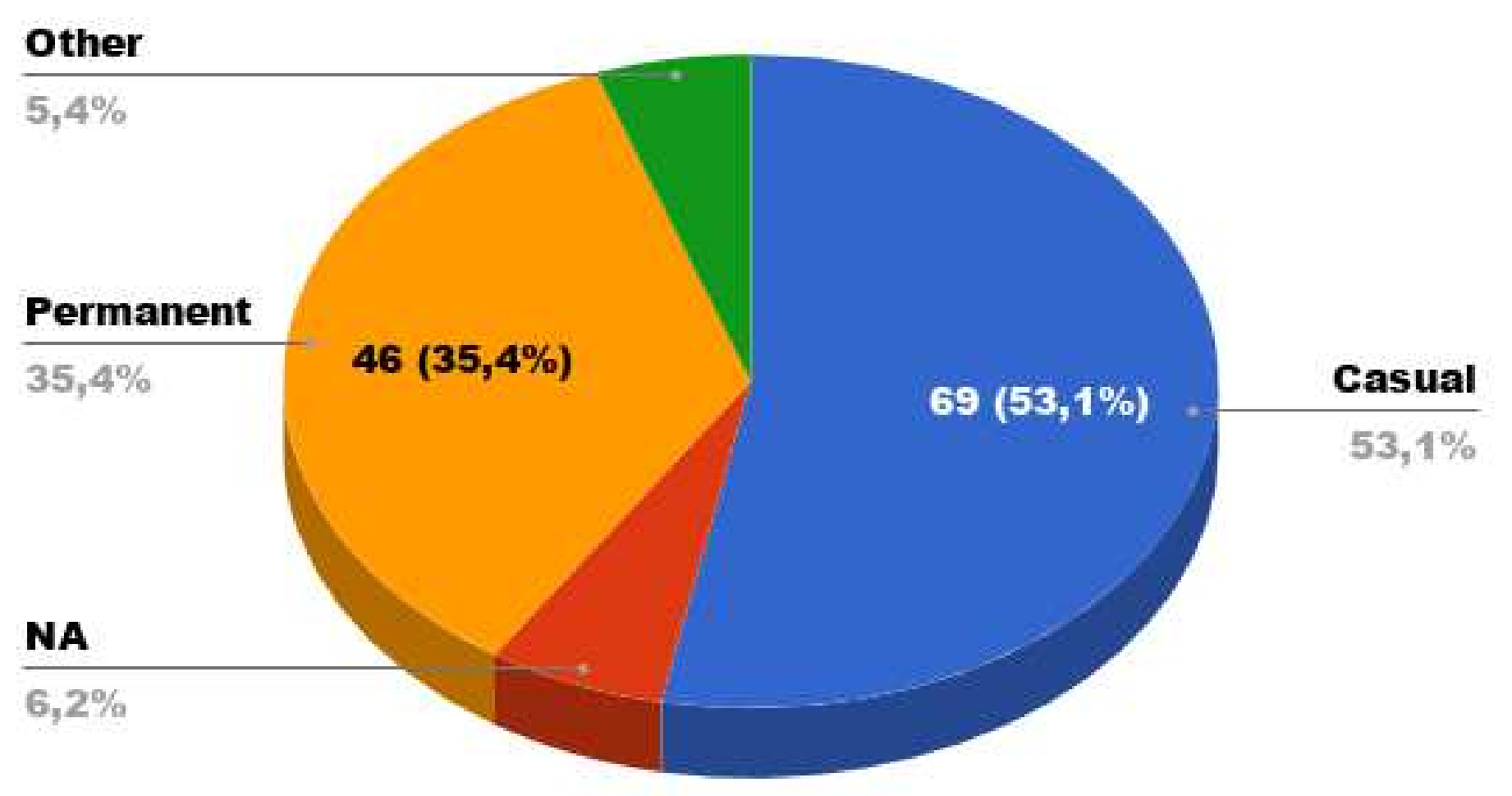
We report the **results from selected Qs** of the survey in order to:

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

The present **results** refer to the responses of **149 participants** (130 completed the survey, 17 opted-out and 2 only suggested other contacts).

D. Accessibility & maintenance issues

Almost half of the resources listed in our survey (**48.5%**) is **barely accessible**. Only **10%** of the resources is **accessible and available**, **3.1%** is **partially accessible**, **36.9%** is **available upon request**, 0.8% is available upon request and only for selected parts (NA's = 0.8%).



The necessity of a national repository is of the highest urgency → consider that most of those owning speech resources in our survey (**about 53%**) fall within the **casual workers** category (e.g. workers without a permanent position nor a permanent affiliation to an institution). Only 35.4% of the respondents declared a permanent position.

E. Ethics & legal issues concerning oral resources

One further information emerging from our survey relates to ethics and legal issues, which are **addressed by the respondents only in 46.2% of the cases**. This has unavoidable effects especially on the accessibility and reusability of such resources and represents something all the subjects involved in the creation and collection of future resources should be aware of.

F. The CLARIN EU infrastructure in our survey's scientific community

Only 31.5% 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. 64.6%) who would agree in storing their archives and their speech resources in national repositories.

This manifestation of interest should give CLARIN's mission more strength and actuality.

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 consciousness, 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**. 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.

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.

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