Building a home for Italian audio archives

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

Niccolò Pretto
Institute of Computational Linguistics, "A. Zampolli"
CNR, Pisa, Italy
niccolo.pretto@ilc.cnr.it

Monica Monachini
Institute of Computational Linguistics, "A. Zampolli"
CNR, Pisa, Italy
monica.monachini@ilc.cnr.it

Maria Francesca Stamuli
Ministry of Culture and Tourism, Florence, Italy
mariafrancesca.stamuli@beniculturali.it

Silvia Bianchi
DSFUCI, Siena University
Arezzo, Italy
bianchi.silvia83@gmail.com

Pierangelo Bonazzoli
Unione dei Comuni Montani del Cosentino, Italy
pierangelobonazzoli@casentino.toscana.it

Abstract

Audio and audiovisual archives are at the crossroads of different fields of knowledge, yet they require common solutions for both their long-term preservation and their description, availability, use and reuse. Archivio Vi.Vo. is an Italian project financed by the Tuscany Region, aiming to (i) explore methods for long-term preservation and secure access to oral sources and (ii) develop an infrastructure under the CLARIN-IT umbrella offering several services for scholars from different domains interested in oral sources. This paper describes the project's infrastructure and its methodology through a case study on the Caterina Bueno's audio archive.

1 Introduction

Audio and audiovisual archives are scattered all over the Italian peninsula, from researchers' private houses, to universities and research centres, from cultural institutions (e.g., Istituti per la Resistenza), to State institutions, such as State Archives and Libraries. The pilot survey made by Galatà and Calamai in 2018 has emphasised the status of precariousness, instability, and insecurity that affects audio and audiovisual archives available at different communities of Italian researchers. Almost half of the resources listed in survey (49.6%) were barely accessible. Only 9.2% of the resources was accessible and available, 4.6% was partially accessible, 35.1% was available upon request, 1.5% is available upon request and only for selected parts. As for the resources which were declared to be accessible, the access policies were as follows: only 9.2% of these resources was freely accessible online (with no authentication); 7.6% was accessible online via authentication; 29% was accessible onsite (i.e. where the resources are physically stored). As for the long-term maintenance and preservation the answer receiving the highest number of responses was nobody (43%), followed by reference Institutes, such as Associations, Foundations, libraries and their archives (17%), reference Universities (16%), the owners/individuals themselves (15%)[1]. Several research projects in recent years aimed to disseminate audio and audiovisual archives, which are collected over the years by both researches and amateur fieldworkers: some examples are, among others, *Grammo-foni. Le soffitte della voce*, also referred as Gra.fo (Calamai and Biliotti, 2017), *Voci, parole e testi della Campania*[2], *I granai della memoria*[3], *Circolo Gianni Bosio Audio Archives*[4]. Nevertheless, fragmentation and lack of common and shared standards are often the common features of certain initiatives, whose duration over time crucially appears to be dependent on the duration of external funding, if any. Moreover, a researcher working with audio archives is not necessarily competent in long-term preservation of audio data and data management. Eventually, not all the research projects

[1] Further details available in (Galatà and Calamai, 2019)

[2] www.archivicampani.unina.it/archivi_campani_dev Last visited August 20th, 2020

[3] www.granaidellamemoria.it Last visited August 20th, 2020

[4] www.circologiannibosio.it/archivio.php Last visited August 20th, 2020

dealing with audio archives receive financing for all the different professional profiles involved in their preservation, managing and valorisation.

Given this picture, it appears urgent to provide an infrastructure offering: 1) a long-term preservation service for audio archives, 2) a shared set of metadata compliant with the main international standards and FAIR principles, and 3) an access interface which takes into account the peculiarities of the audio modality and which is able to support researchers in different disciplines. This paper presents how the Archivio Vi.Vo. project tackles these problems, illustrating the overall adopted methodology and the developed infrastructure.

2 The Archivio Vi.Vo. project

2.1 The project

In 2019, Regione Toscana decided to support the project Archivio Vi.Vo., which aims to catalogue and disseminate oral archives. Partners involved are: Siena University (Silvia Calamai), CNR-ILC & CLARIN-IT (Monica Monachini), Soprintendenza Archivistica e Bibliografica della Toscana (Maria Francesca Stamuli) and Unione dei Comuni del Casentino (Pierangelo Bonazzoli). In order to reach the above ambitious objectives, Archivio Vi.Vo. concentrates most of the efforts on the design and development of an architecture, hosted by CLARIN-IT, the Italian node of the CLARIN research infrastructure, that could be used by several other projects concerning audio archives. A crucial step towards this main aim concerns the definition of the metadata set(s) used to describe the data. The set has to be compliant with international archival standards, such as with ISAD(G) and ISAAR, as well as several others derived from different disciplines (cf. below Section 3.1).

2.2 The case study

The architecture (cf. Section 3) is in the process of being validated on a specific audio archive, namely the Caterina Bueno's audio archive, which appears to be rather challenging, for the following reasons: (i) it has a complex archival history, (ii) it is in a very poor conservation condition, and (iii) it contains highly heterogeneous audio material.

Caterina Bueno (San Domenico di Fiesole, IT, 2nd April 1943 – Florence, IT, 16th July 2007) was an Italian ethnomusicologist and singer. Her work as a researcher has been highly appreciated for its cultural value, as it allowed the collection of many Tuscan and central Italy's folk songs that had been orally passed down from one generation to the next until the 20th century (when this century old tradition started to vanish). Her work as a singer was always oriented towards research. At the age of twenty, she started travelling through the Tuscan countryside and villages recording Tuscan peasants, artisans, common men and women singing any kind of folk songs: lullabies, *ottave* (rhyming stanzas sung during improvised contrasts between poets), *stornelli* (monostrophic songs), narrative songs, social and political songs, and much more. These were the same songs that she sang in her performances, making them well-known and appreciated both in Italy and abroad in the second half of the 20th century, when she was at the pinnacle of her career. Caterina Bueno's sound archive is composed of about 476 analogue carriers (audio open-reels tapes and compact cassettes), corresponding to more than 700 hours of recording, and it was digitised during the PAR-FAS project Gra.fo. The analogue recordings were located at two different owners: part of them were stored at Caterina's heirs' house, while the rest was kept by the former culture counsellor of the Italian Municipality of San Marcello Pistoiese, in the Montagna Pistoiese, where a multimedia library was supposed to be set up. Unfortunately, disagreements and misunderstandings between the two parties have so far made the archive fragmented and inaccessible to the community. Both owners, independently, have turned to Silvia Calamai for the reassembly of the whole archive in the digital domain, in respect of the artist's wishes. After being digitised, the carriers were returned to their owners.

In several cases, the original carriers were devoid of all the contextual information (place and date of recordings, speakers involved in the recordings). In other cases, the open-reel tapes were recorded at different speeds and using different track head configurations, thus making rather complex the digitisation process and the creation of access copies. From this respect, Caterina Bueno's audio archive represents

an extreme test case where different levels of complexity call into question different professional profiles and skills.

3 Building the home for audio archives

3.1 Data and metadata

In order to preserve and provide access to analogue audio recordings (e.g., the compact cassette or the open reels), it is essential to digitise them. The result of the digitisation process is the digital *preservation copy*, which is composed by the audio content as well as several other information about the carrier (such as the photo of the carrier itself, or its box)[5]. As the name suggests, the preservation copy is the "means" for safeguarding the content of the audio documents and it can be considered as the new digital master for long-term preservation. Another important concept to be defined is the *archival unit*. In audio and audiovisual archives, it is defined as a set of data and documents pertaining to the very same communicative event, per unit of time and place. Archival unit is the outcome of a meticulous process involving listening, analysis and comparison. Sometimes audio content needs to be re-organised. For example, an archival unit could be composed by content that is stored in several physical carriers (and, therefore, in several preservation copies), or vice versa, several archival units could be stored in the same physical carrier[6]. Given the absence of a one-to-one relationship between the physical carrier (i.e., compact-cassettes, open-reels) and the archival unit, the preservation copies are kept separately from the archival units (Mulè, 2003; Calamai et al., 2014; Stamuli, 2020).

This approach leads to a very complex set of metadata, articulated along three different layers: (i) metadata for the description of the preservation copy, (ii) metadata for the description and managing of oral sources as items of an (audio) archive (archival unit), and (iii) metadata expressing the relationship between the preservation copy metadata and the digital archive metadata.

In Archivio Vi.Vo., a customised set of metadata has been defined for (i), inspired by other international standards for audio material description, in particular the one proposed by Association of Sound and Audiovisual Archives (IASA Technical Committee, 2009). The project adopted ISAD(G) and ISAAR standards for the archival units (ii), encoding the information about archival material with Encoded Archival Description (EAD) and Encoded Archival Context (EAC) standard data models. One of the main challenges is to make these metadata structures interoperable with the CLARIN VLO infrastructure component which is part of CLARIN's Component Metadata Infrastructure and can cope with many different metadata descriptions, as long as they are implemented through (or converted to) the Component Metadata framework. The metadata structure for expressing the relationship between the preservation copy and the archival unit (iii) is based on the methodology described below.

3.2 From preservation copies to archival units

The methodology formalised and adopted in Archivio Vi.Vo. is composed by several steps. All the operations performed during these steps and the information inserted by audio technicians, researchers and/or cataloguers are stored and duly described by a set of appropriate metadata, thus maintaining the relation between preservation copies and archival units. The methodology starts with the creation of the preservation copy of a single audio recordings. This phase has proved to be very delicate and time-consuming.

Sometimes the audio recording is not easily accessible, due to, e.g., different speeds, configurations or digitisation errors (Pretto et al., 2020). In these cases, if necessary, researchers or audio technicians recur to the concept of clip[7] in order to separate parts with different speeds, channels with different recordings or recordings in different directions. In Archivio Vi.Vo. a clip is defined as a duplicate of an audio segment extracted from a preservation copy. One or more clips can be extracted from a preservation copy. In some cases, the clips are the result of a restoration operation, necessary for the use of the sound content. The process of creating (and restoring) the clips must in no way modify the preservation copy.

[5] An extended description of the preservation copy is available in (Bressan and Canazza, 2013)

[6] Several other kinds of transformation could be performed, but their description goes beyond the scope of the paper

[7] The concept of clip is commonly used to indicate data of either video or audio that has been clipped out (copied) from a larger environment such as a reel or a video tape

The resulting clips will be correctly accessible and allow the researcher/cataloguer to listen, analyse and describe their contents. In case some parts of the very same clip belong to different events (and, therefore, to different archival units, see Section 3.1), they will be segmented accordingly and new sub-clips will be created (archival unit clips). In some cases, some archival unit clips, derived from different preservation copies, would be part of a same event, therefore, they will also be part of the same archival unit. As soon as all the archival unit clips will be ordered, and all the missing metadata required by ISAD(G) will be added, the archival unit will be created and available through the access interface.

3.3 The infrastructure

As for the infrastructure, ILC4CLARIN and the CLARIN-IT national data centre (Monachini and Frontini, 2016), will implement new experimental approaches to preservation, management and access to audio data and metadata. The experimental activity aims to adopt the model and the high-performance computing and archiving services of the new GARR network infrastructure, built along the Cloud paradigm[8]. The project will also exploit the federated identity service of the CLARIN infrastructure, in order to manage users' access. A robust system for managing authentication is essential for audio and audiovisual archives because of the frequent privacy, ownership, and copyright issues concerning their content (Kelli et al., 2019; Kelli et al., 2020). Several classes of users are considered each of them with different access grants.
The infrastructure consists of two different parts. The first one provides a data and metadata entry interface for archivists, archive owners or, in general, researchers who want to preserve the legacy. The system is highly complex: it must be able to manage several international standards and several kinds of specific functionalities. Considering the complexity of the project, the infrastructure would difficulty be developed from scratch. Therefore, as a first step, ten archival software were evaluated on the base of several features and technologies (standards, programming languages, frameworks, DBMS, license, etc.). The selected software was the open-source software xDams[9]. Three main characteristics influenced the adoption of the software: (i) the completeness of its standards coverage, (ii) its extensible no-sql database as well as (iii) the open-source license. The second part of the infrastructure consists of an access interface able to support researchers of different disciplines in discovering and studying audio or audiovisual documents. In order to study the interaction with the software, two mockups were developed for studying and testing the interfaces for inserting and cataloguing the digitised documents (Figure 1a) and for accessing their content (Figure 1b), respectively. The two mockups have been developed with the frameworks Vue.js and Bootstrap, respectively, Web Audio API, as well as Peak.js and Audiowaveform, two libraries developed by BBC[10].

4 Final remarks

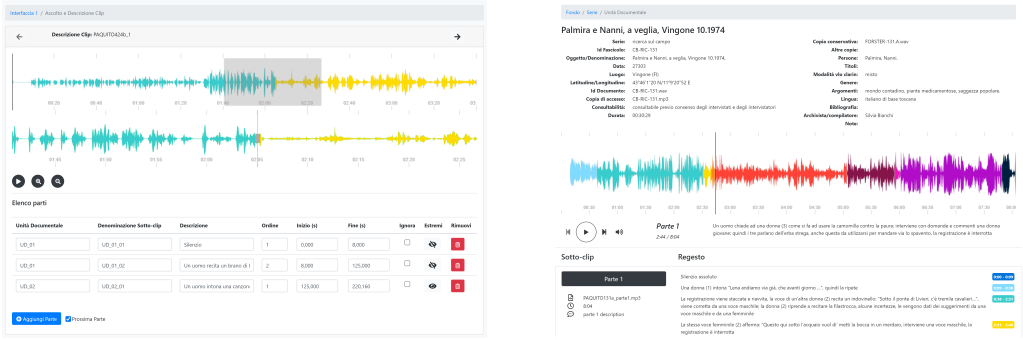
Archivio Vi.Vo. constitutes a pilot case study within CLARIN-IT to experiment with methods for long-term preservation and secure access to oral source and offer targeted services for both specialists and the general public interested in these data. Archivio Vi.Vo. aims not only to develop an infrastructure for preservation, description and use of audio archives, but, more ambitiously, to define and develop a model for the management, protection and enhancement of archival heritage which can be replicated, even outside the context of Tuscany.

References

Federica Bressan and Sergio Canazza. 2013. A systemic approach to the preservation of audio documents: Methodology and software tools. *Journal of Electrical and Computer Engineering*, 2013:21 pages.

Silvia Calamai and Francesca Biliotti. 2017. The gra.fo project: from collection to dissemination. *Umanistica Digitale*.

[8] `cloud.garr.it` Last visited August 20th, 2020
[9] `www.xdams.org` Last visited August 20th, 2020
[10] `github.com/bbc/peaks.js` and `github.com/bbc/audiowaveform` Last visited August 20th, 2020



(a) (b)

Figure 1: (a) section of the interface for creating archival units from preservation copies: the clips derived from the duplicate of preservation copy are segmented, described, ordered and assigned to an archival unit; (b) the interface for accessing the archival units contents

Silvia Calamai, Francesca Biliotti, and Pier Marco Bertinetto. 2014. Fuzzy archives. what kind of an object is the documental unit of oral archives? In Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen, and Ewald Quak, editors, *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, pages 777–785, Cham. Springer International Publishing.

Vincenzo Galatà and Silvia Calamai. 2019. Looking for hidden speech archives in italian institutions. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, number 159, pages 46–55. Linköping University Electronic Press.

IASA Technical Committee. 2009. *Guidelines in the Production and Preservation of Digital Audio Objects: standards, recommended practices, and strategies. IASA TC-04*. International Association of Sound and Audiovisual Archives, 2nd edition.

Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavriilidou, Pavel Stranák, et al. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, Italy, 8-10 October 2018*. Linköping University Electronic Press.

Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Straňák, et al. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. *Selected Papers from the CLARIN Annual Conference 2019, Leipzig, Germany, 30 September, 2019*.

Monica Monachini and Francesca Frontini. 2016. CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics*, 2(2):11–30.

Antonella Mulè. 2003. Le fonti orali in archivio. un approccio archivistico alle fonti orali. *Archivi per la storia*, 16(1):111–129.

Niccolò Pretto, Alessandro Russo, Federica Bressan, Valentina Burini, Antonio Rodà, and Sergio Canazza. 2020. Active preservation of analogue audio documents: A summary of the last seven years of digitization at csc. In *Proceedings of the 17th Sound and Music Computing Conference, SMC20, Torino*, Torino.

Maria Francesca Stamuli. 2020. Fonti orali, documenti e archivi: riflessioni e proposte per la nascita di un 'archivio vivo'. In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Studi AISV 6*.