Balancing FAIR and GDPR: a governance framework for oral archives

Elvira Mercatanti°, Monica Monachini°, Giovanni Abete', Silvia Calamai", Sergio Canazza°°, Alessandro Casellato*, Virginia Niri", Cesarina Vecchia', Giulia Zitelli Conti*, Giada Zuccolo°°

CNR-ILC, Pisa°, Università degli Studi di Napoli Federico II', Università degli Studi di Siena"
Università degli Studi di Padova°°, Università Ca' Foscari Venezia*
elviramercatanti@cnr.it, monica.monachini@cnr.it, giovanni.abete@unina.it,silvia.calamai@unisi.it,
canazza@unipd.it, casellat@unive.it, virginia.niri@unisi.it, cesarina.vecchia@unina.it,
giulia.zitelliconti@unive.it, giada.zuccolo@studenti.unipd.it

Abstract

This paper presents a governance framework developed within the research project ROADS to support the sustainable management of oral archives, which constitute essential linguistic resources for interdisciplinary research and cultural heritage preservation. Oral archives raise complex ethical and legal challenges due to the hybrid nature of voice data, which function simultaneously as historical documents, scientific sources and biometric identifiers, thereby creating tensions between open science principles and data protection regulations. The proposed framework integrates FAIR principles (Findable, Accessible, Interoperable, Reusable) with Privacy by Design and the GDPR accountability principle through a multilayered approach. It introduces an access model that distinguishes between publicly available metadata and controlled access to identifiable audio materials, following trusted repository standards. The framework also incorporates consent management procedures and safeguards for legacy collections, enabling responsible data sharing while preserving scientific usability. More broadly, ROADS provides a transferable model to guide the transition from project-based archives to FAIR, sustainable and reusable research resources, ensuring compliance with data protection requirements and respect for the sensitivity of the documented contexts.

Keywords: GDPR, speech data, ethics, oral archives, FAIR principles, language resources

1. Introduction

Access to speech data is essential for linguistic research, language technologies and cultural heritage preservation (De Dominicis, 2002; Sornicola et al., 2019). A central challenge in oral archives lies in the hybrid nature of voice data: audio recordings simultaneously function as historical documents, scientific sources and biometric identifiers, creating tensions between open science principles and data protection regulations.

Managing oral archives requires an interdisciplinary approach integrating humanistic, archival, technological and legal expertise. Researchers must reconcile data openness and reuse with the protection of data-subject rights while ensuring long-term sustainability and integration into Digital Humanities ecosystems. In Italy, the digitization and reuse of oral archives remain limited due to heterogeneous practices, local preservation constraints and regulatory uncertainty. Recent regulatory frameworks, particularly the General Data Protection Regulation (GDPR), have reshaped how sensitive linguistic resources can be collected, processed and shared. Ethical principles such as fairness, transparency and trust are no longer abstract ideals but concrete design constraints for research infrastructures.

Within this context, the ROADS project has developed a governance framework to overcome the fragmentation of Italy's oral heritage through standardized scientific and regulatory practices. To guarantee the long-term sustainability of the model, ROADS relied on legal experts embedded within the participating institutions-professionals capable of mediating between the philosophy of Open Science and the stringent constraints imposed by the GDPR. Their contribution has been crucial in ensuring the legal integrity of the resources produced, making them fully legitimate and transparent scientific sources for future generations (Abete et al., 2026).

The paper is structured as follows: Section 2 describes ROADS and its pilot archive, together with the challenges encountered; Section 3 presents the proposed solutions for FAIR- and GDPR-compliant reuse; Section 4 outlines the conclusions.

2. The ROADS project as a model for oral data

ROADS[1] is a national Italian Project of Relevant National Interest designed to coordinate and sustain Italy's oral heritage by developing models and tools for the recovery, preservation, description and scholarly reuse of oral archives, adopting a FAIR-by-design approach integrating technical, ethical

[1] https://csc.dei.unipd.it/roads-project/index.html

Proceedings of LEGAL2026 and CALD-pseudo 2026 @ LREC 2026, pages 81–85
12 May 2026.

and legal considerations from the beginning of the data lifecycle (Abete et al., 2025b). Core project activities include a national survey of oral archives in Italian public universities, development of a management and access infrastructure, deposit of a pilot archive and training initiatives to build capacity for sustainable stewardship of oral sources (Abete et al., 2025a).

ROADS addresses two types of sources: (i) pre-existing historical oral archives, collected before the current regulatory framework; (ii) new oral interviews collected within the project, targeting a representative selection of researchers to gather information on the genesis of collections, their size and composition, archival criteria, and conservation status. This dual perspective enables differentiated methodological and legal solutions tailored to distinct data-production contexts, helping make FAIR principles operational (Calamai and Frontini, 2018; Wilkinson et al., 2018) even when normative and ethical constraints are heterogeneous.

A key element of the governance framework is the Data Management Plan (DMP), which specifies how data are collected, documented, preserved and shared. The DMP applies FAIR principles across all stages of the data lifecycle and links them to legal, ethical and technical requirements. It defines procedures for consent, anonymisation, licensing and access control, ensuring GDPR compliance and safeguarding sensitive historical and biometric information. This ensures that resources are usable for research while respecting the rights and privacy of data subjects.

2.1. The pilot archive

The selected pilot archive is the research collection of historian Gabriella Gribaudi[2], based on fieldwork conducted since 1974 on the social history of Southern Italy during World War II (Gribaudi, 1980, 1990, 2005, 2016, 2023). The collection comprises 148 audio carriers (audiocassettes, minicassettes, and digital audio tapes) and approximately 189 interviewees (born 1909-1945), including both direct witnesses and individuals who reported family narratives. Beyond its historical relevance, the archive is particularly suitable for linguistic analyses, as it contains rich, naturally produced speech with speaker and contextual diversity supporting phonetic, sociolinguistic, and discourse-oriented studies. This case provides a realistic benchmark for implementing differentiated access and reuse policies for historically sensitive, inherently identifiable oral data.

[2]Professor at the University of Naples Federico II and founder and first president of the Italian Oral History Association.

2.2. Challenges requiring legal and governance measures

Oral archives in the Italian context are often preserved locally, described with heterogeneous metadata practices, and shared under unclear conditions. This hinders discoverability, long-term sustainability, and reuse. A major source of complexity lies in the intrinsic nature of oral data: audio recordings contain not only biographical information and opinions, but also biometric traits (voice) and dialectal or cultural cues.

Within the project, a further limitation arises from the coexistence of legacy collections and newly collected data. Legacy recordings, such as the Gribaudi archive, were produced before the GDPR and often lack documentation that would now be expected (e.g., explicit consent forms, clear information on intended dissemination, and standardized provenance). In many cases, data subjects cannot be contacted due to decease or irretrievability, which makes it impossible to update consent and requires careful legal framing and proportionate safeguards. By contrast, newly collected interviews can be designed to meet GDPR transparency and accountability requirements from the outset, but they still contain inherently identifiable voice data and potentially sensitive contextual information. These two regimes create a practical governance challenge: a single infrastructure must support FAIRness and scientific usability while applying differentiated access and reuse rules aligned with the origin, documentation, and sensitivity of each dataset.

3. Solutions for FAIRness, legal and ethical compliance

Managing personal data within ROADS required fulfilling regulatory obligations to ensure legitimate research activity. These steps constitute the backbone of a protection system that safeguards individuals. In line with the accountability principle, ROADS formalized roles among partners, identified legal bases, defined secure preservation protocols, and implemented transparent information flows to data subjects.

3.1. Legal framing and governance

The project follows two parallel methodological tracks: the ethical valorization of historical oral archives and the collection of new testimonies, both aimed at building a sustainable and secure research infrastructure compliant with FAIR principles. Given the multicentric and interdisciplinary nature of the project, governance has been formalized through a joint controllership agreement

(Art. 26 GDPR) among ROADS partners. This instrument clarifies each partner's responsibilities, appoints a single contact point for exercising data-subject rights (Arts. 15–22 GDPR), and identifies the national research infrastructure CLARIN-IT as the technological custodian supporting long-term preservation.

One of the main challenges is that informed consent cannot be obtained for part of the legacy materials because many interviewees are deceased or no longer traceable. ROADS addresses this issue through a formal "Diligent Search" protocol, based on public notices on institutional websites that inform data subjects (or their heirs) about digitization and reuse while preserving the right to object on legitimate grounds. Although the GDPR does not apply to deceased people, this approach is aligned with Art. 2-terdecies of the Italian Privacy Code, which enables heirs to exercise the data subject's rights post mortem. The framework further balances the public interest in protecting cultural heritage with privacy safeguards by minimizing or removing identifying metadata and restricting access to full audio through controlled procedures (e.g., excerpts or partial access), thereby reducing the risk of unlawful or inappropriate reuse.

3.2. A multilevel system of legal bases

ROADS relies on an integrated set of legal bases, calibrated to the nature of the processing and the institutional mandate of the partners:

- Scientific research (Art. 6(1)(e) GDPR): core research activities are grounded in the public interest. This legal basis derives from a combined reading of the GDPR and Art. 2-ter of the Italian Privacy Code (D.Lgs. 196/2003), which recognizes scientific and historical research as a primary institutional function of Universities and Public Research Bodies. This framework ensures that processing is not contingent upon individual withdrawal where the data serves a broader collective scientific purpose, provided that the principle of data minimization is strictly observed.

- Archiving and Historical Research (Art. 9(2)(j) GDPR; Art. 2-sexies Italian Privacy Code): this legal basis is particularly relevant for legacy collections where consent cannot be obtained due to the decease of the data subject or their untraceability. In such cases, the framework incorporates Art. 2-terdecies of the Italian Privacy Code, which provides that heirs may exercise the relevant data protection rights on behalf of the deceased. Appropriate safeguards are implemented through a formal "Diligent Search" protocol (e.g., public notices and the right to object), combined with robust technical and organizational measures, including controlled access via ILC4CLARIN. In particular, the protocol requires the anonymization of descriptive metadata in legacy archives, ensuring that identifying references are protected and not publicly accessible, while restricting full audio access to authenticated researchers in order to mitigate risks such as unauthorized voice harvesting.

- Informed Consent and releases (Art. 6(1)(a) GDPR): in coordination with applicable copyright provisions, this legal basis governs optional and ancillary processing activities. Adopting a granular approach, ROADS distinguishes between essential research operations and specific authorizations—such as video dissemination, third-party reuse, or potential commercial exploitation—which remain fully revocable by the data subject at any time.

To illustrate this model, consider the ILC4CLARIN repository workflow for video interviews. The system adopts a differentiated access model: a curated "Partial Version" is made available for public dissemination, while the complete recording is distributed under a "Restricted" license. Access to the full version requires institutional Single Sign-On (e.g., IDEM/Edugain), thereby preventing unauthorized voice harvesting and ensuring traceability in line with GDPR accountability requirements. For legacy data lacking valid consent, the "Diligent Search" protocol is systematically applied.

3.3. Operational implementation: acknowledgement, stratified consent and transparency

The main complexity lies in the intrinsic nature of oral data: audio recordings may contain not only biographical information and opinions, but also biometric traits (voice) and dialectal or cultural cues. This requires a granular governance approach that distinguishes what is necessary for science from what pertains to dissemination. To translate this complexity into practice, the project implements a stratified consent architecture (Modules 01, 02, 03) that supports informed control:

MOD 01 - participation and originality: formalizes participation and a declaration that provided content does not infringe third-party rights.

MOD 02-A - mandatory acknowledgement of the information notice: supports scientific transparency and accountability, including awareness that identifying metadata may be publicly available for scholarly attribution and long-term findability.

MOD 02-B and MOD 03 - optional modules: separate choices on re-contact, dissemination-oriented

video use, and third-party reuse for external scientific/didactic purposes.

3.4. Security, preservation and scientific authorship

In accordance with what is stated in the information notice, and following the participant's acknowledgment of the FAIR protocols, the security of data processing is ensured through a multi-level, accredited access system designed to balance individual protection with the needs of the scientific community:

- Attribution and traceability: since these are methodological contributions provided by scholars, the identifying metadata (name and affiliation) are made public. This step is essential not only to meet FAIR requirements, but also to ensure the proper scientific authorship of the collected testimony and its traceability over time. The partial version of the archives, comprising descriptive metadata and curated audio excerpts, is available in Open Access to the general public. This version is indexed by general search engines to ensure maximum findability, but it is carefully processed to reduce the risk of re-identification.

- Multi-level and accredited access: while attribution is public, the full audio/video files are deposited in a restricted-access digital archive. Access is protected by authentication protocols and reserved exclusively for the international scholarly community, preventing indiscriminate dissemination of the content and of biometric data beyond the research context. Access to these integral versions is strictly shielded: they are not indexed by search engines and are accessible only to verified scholars through strong authentication protocols (e.g., IDEM/Edugain/Shibboleth) via the ILC4CLARIN certified repository. This prevents indiscriminate dissemination and automated voice-harvesting, limiting use to verified research contexts.

- Preservation: the data are deposited in the certified repository, which ensures protection against unauthorized access and the safeguarding of the information assets beyond the duration of the project, in accordance with Article 99 of the Privacy Code for purposes of public interest. Specifically, data processed for historical or scientific purposes are preserved indefinitely as a public information asset, provided they are not used for decisions affecting the specific data subject.

4. Conclusions

ROADS demonstrates that managing oral archives is a complex interdisciplinary challenge, combining legal compliance, ethical accountability and robust infrastructure. Challenges addressed include the hybrid nature of voice data, which simultaneously function as historical testimony, scientific sources and sensitive biometric identifiers, and the temporal stratification of archives, requiring harmonization of legacy pre-GDPR collections with current European standards.

The approach preserves the integrity of recordings in the transition from analog to digital formats and defines specific protocols to protect the privacy of witnesses, often deceased or unreachable. Structured governance, multilevel access, and rigorous transparency protocols make ROADS a reference model for Digital Humanities, ensuring historical memory is preserved within a fully legal and ethically robust ecosystem.

From a research-infrastructure perspective, the project operationalizes FAIR-by-design for inherently identifiable speech: public metadata enable discovery and attribution, while controlled access preserves scientific usability and data-subject rights. Components such as role allocation, joint controllership, calibrated legal bases, access tiers, and documentation practices can be replicated in other projects dealing with sensitive speech or oral history collections. Currently, the application and replicability of the defined model are being tested on project archives to support their transition into FAIR, sustainable and GDPR-compliant resources.

5. Acknowledgments

This work is supported by the PRIN PNRR 2022 project ROADS (MUR P20229S48H), by CLARIN-IT, the Italian node of the CLARIN ERIC research infrastructure, and by the H2IOSC Project-Humanities and Cultural Heritage Italian Open Science Cloud, funded by the European Union-NextGenerationEU (NRRP M4C2, project code IR0000029;see the project website at https://www.h2iosc.cnr.it/). We gratefully acknowledge Rosaria Deluca, Responsible for the Privacy Service of the CNR Research Area in Pisa, for her expert legal guidance and support of the ROADS project. Her advice and assistance were indispensable to our work.

6. Bibliographical References

Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Elvira Mercatanti, and Monica Monachini. 2026. La filiera legale di ROADS. Una proposta FAIR per archivi orali analogici.

Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Elvira Mercatanti, Monica Monachini, Virginia Niri, Cesarina Vecchia, Giulia Zitelli Conti, and Giada Zuccolo. 2025a. Fare didattica con gli archivi orali: la fonte orale al crocevia di discipline e di saperi.

Giovanni Abete, Cesarina Vecchia, Silvia Calamai, Alessandro Casellato, Sergio Canazza, Elvira Mercatanti, Monica Monachini, Roberta Ottaviani, Giulia Zitelli Conti, and Giada Zuccolo. 2025b. On the lifecycle of Italian oral archives: the ROADS project. In *La voce della grammatica. Nuove prospettive sull'interazione tra fonetica e morfologia, sintassi, lessico*, Università degli Studi di Urbino Carlo Bo. Associazione Italiana di Scienze della voce.

Silvia Calamai and Francesca Frontini. 2018. FAIR data principles and their application to speech and oral archives. *Journal of New Music Research*, (47):339–354.

Amedeo De Dominicis, editor. 2002. *La voce come bene culturale*. Carocci, Roma.

Gabriella Gribaudi. 1980. *Mediatori: antropologia del potere democristiano nel Mezzogiorno*. Rosenberg & Sellier.

Gabriella Gribaudi. 1990. *A Eboli. Il mondo meridionale in cent'anni di trasformazione*. Marsilio, Venezia.

Gabriella Gribaudi. 2005. *Guerra totale. Tra bombe alleate e violenze naziste. Napoli e il fronte meridionale 1940-1944.* Bollati Boringhieri, Torino.

Gabriella Gribaudi. 2016. *Combattenti, sbandati, prigionieri. Esperienze e memorie di reduci della seconda guerra mondiale*. Donzelli, Roma.

Gabriella Gribaudi, editor. 2023. *Terra bruciata. Le stragi naziste sul fronte meridionale*. Guida, Napoli.

Rosanna Sornicola, Giovanni Abete, Elisa D'Argenio, and Cesarina Vecchia. 2019. Raccontare un archivio di fonti orali: il progetto Voci, parole e testi della Campania. In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale*, volume 6 of *Studi AISV*, pages 75–93. Officinaventuno, IT.

Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5:180118.