The Gra.fo project: from collection to dissemination

Silvia Calamai e Francesca Biliotti

Dipartimento di Scienze della Formazione, Scienze Umane e della Comunicazione Interculturale,
Università di Siena
`silvia.calamai@unisi.it`, `francesca.biliotti@unisi.it`

Abstract. Oral archives collected by professional scholars and ordinary people interested in dialects and ethnology are a precious resource for various fields of study (from linguistics to anthropology, from economy to history and politics, etc.) and may contain documents that could be labeled as products of intangible cultural heritage, thus deserving safeguard. *Grammo-foni. Le soffitte della voce* (*Gra.fo*), a two-year project jointly conducted by Scuola Normale Superiore and the University of Siena (Regione Toscana PAR FAS 2007-13), discovered, digitized, cataloged and disseminated via a web portal nearly 3000 hours of speech recordings stemming from around 30 oral archives collected by scholars and amateurs in the Tuscan territory. Having preserved such a significant collection of oral documents (e.g. oral biographies, ethno-texts, linguistic questionnaires, oral literature), *Gra.fo* constitutes a precious repository of Tuscan memory and provides a first-hand documentation of Tuscan language varieties from the early 1960s to the present day [14], [12]. In this article, the *Gra.fo* project will be described in all its stages, which involve: fostering the level of awareness on the importance of preserving this valuable cultural heritage product; contacting the oral recordings' owners and co-signing legal agreements for the temporary borrowing of the recordings and accompanying materials; collecting and digitizing the recordings and the accompanying materials; cataloging (with the self-developed software *Audio-grafo*) and partially transcribing the oral documents; implementing the downloadable online catalog <`http://grafo.sns.it/`>, an open-ended repository of oral texts which have hitherto been known to a very limited number of potential users. Some problematic issues related to the treatment of oral archives will also be discussed, together with the proposed solutions. These concern the carrier/document relation, the treatment of confidential information, and the cataloging of documents within other documents.

Abstract. Gli archivi orali raccolti da studiosi e appassionati di etnologia e dialetti sono una risorsa preziosa per molte discipline (dalla linguistica all'antropologia, dalla sociologia alla storia e alla politica ecc.) e possono contenere al loro interno documenti etichettabili come "beni culturali immateriali", e come tali degni di tutela. *Grammo-foni. Le soffitte della voce* (*Gra.fo*), un progetto biennale condotto congiuntamente dalla Scuola Normale Superiore e dall'Università di Siena (Regione Toscana PAR FAS 2007-13), ha scoperto, digitalizzato, catalogato e reso fruibili attraverso un portale online quasi tremila ore di parlato provenienti da circa trenta archivi orali raccolti

nel territorio toscano da studiosi e appassionati. Grazie alla vastità e all'eterogeneità dei documenti preservati (biografie orali, etnotesti, questionari linguistici, letteratura orale), *Gra.fo* costituisce un archivio preziosissimo della memoria toscana e restituisce una documentazione di prima mano delle varietà linguistiche toscane fin dai primi anni '60 [14], [12]. L'articolo descrive il progetto *Gra.fo* in tutte le sue fasi, suddivise nelle seguenti azioni: promuovere la consapevolezza dell'importanza della conservazione degli archivi orali come beni culturali; contattare i possessori/detentori di archivi orali e stipulare accordi legali per il prestito temporaneo, la digitalizzazione e la diffusione delle registrazioni e dei materiali di corredo; raccogliere e digitalizzare le registrazioni e i materiali di corredo; catalogare (attraverso il software autoprodotto *Audiografo*) e trascrivere (in parte) i documenti orali; costruire un portale online (<`http://grafo.sns.it/`>) per rendere fruibili documenti orali che prima erano conosciuti ad un pubblico molto limitato. Saranno affrontate anche alcune questioni problematiche relative al trattamento dei documenti orali: il rapporto fra supporto e documento, il trattamento delle informazioni sensibili e la catalogazione di documenti contenuti in altri documenti.

1 Introduction

Oral traditions and expressions, including language as a vehicle of intangible cultural heritage; performing arts; social practices, rituals and festive events; knowledge and practices concerning nature and the universe; and traditional craftsmanship are the materials that The UNESCO *Convention for the Safeguarding of the Intangible Cultural Heritage,* Article 2 [27] include under the label "Intangible Cultural Heritage domains". Accordingly, the contents of several oral archives have become part of Intangible Cultural Heritage to be safeguarded. However, this type of safeguarding proves to be problematic because of various issues concerning oral material conservation and accessibility. As far as conservation is concerned, the deterioration of the carriers and the obsolescence of the recording systems make it very difficult to play a recording collected some decades ago [15]. As for oral archive accessibility, private archives are often only known and accessible to the researcher(s) who collected them, while public archives suffer out of lack of communication between different academic fields [7].

The project *Grammo-foni. Le soffitte della voce* (*Gra.fo*), jointly conducted by Scuola Normale Superiore of Pisa and the University of Siena, and financed by the Regione Toscana (PAR FAS 2007-13), addressed these very issues. *Gra.fo* detected and preserved oral documents (e.g. biographies, ethno-texts, linguistic questionnaires, oral literature, etc.) collected on the Tuscan territory and made them available to the public in an online archive.

The creation of an archive incorporating the main oral archives of the region involved different, interconnected stages of work. It was necessary to lay the foundations for an interdisciplinary dialog between linguistics, anthropology, informatics, and archival science, following the example of the work done by the *Maisons des sciences de l'homme* and the *Association française des détenteurs de documents audiovisuels et sonores.*

In the following sections, the different stages of the project will be described: Section 2 deals with the preliminary stages (census, collection and digitization); Section 3 is about cataloging; Section 4 concerns transcription; Section 5 describes the creation of the online archive. The last section (Section 6) will be devoted to the discussion of some problematic issues (together with the proposed solutions), namely the relation between the document and its carrier(s), the ethical and legal issues related to the treatment of confidential data, and the cataloging of documents contained within other documents.

2 The preliminary stages: census, collection and digitization

2.1 Creating a census of Tuscan oral archives

Besides being rich in paper documents [23], Tuscany is also a privileged area for working on oral documents, as it abounds with both public and private sound archives, collected by scholars as well as amateurs. In creating a census of Tuscan oral archives, already existing censuses (namely [2], [4]and [3]) were used and integrated with information about oral archives collected for linguistic and dialectological research purposes, such as *Carta dei Dialetti Italiani*, *Atlante Lessicale Toscano* and *Vocabolario del Fiorentino Contemporaneo*.

Subsequently, a priority list was defined according to three main criteria:

– relevance and antiquity of the materials (older materials might witness dead or disappearing language varieties);
– state of preservation of the materials (priority was given to those materials which looked more damaged and whose content, therefore, was more likely to be lost in the near future);
– geographic representativeness (so that every area of Tuscany was represented in the archive).

2.2 Collecting the materials

Following the above-mentioned priority list, the sound archives' owners were directly contacted and the aims and organization of the project were explained to them. The *Gra.fo* staff met those who accepted to join the project, in order to collect their archives and sign legal agreements for the temporary borrowing and the dissemination of their materials.

In addition, the owners of the archives with no proper bibliography or accompanying material were interviewed so that they could explain the motivation and aims of their research. Indeed, unlike other kinds of materials, oral documents are often obscure objects: usually, the motivation behind them is clear only to the researcher(s) who collected them. Such interviews, called 'Say something about your archive', are crucial as they provide catalogers with the key for interpreting and describing the archive, and the users with an appropriate guide to understand it. The 'Say something about your archive' interview comprises the following questions:

1. How was your archive born?
2. What were the research aims?
3. Which difficulties did you find during your research? In which conditions did you work? How did you find the speakers for your research?

4. Did you publish something out of this research? Do you have any transcription of the material?

5. Was your research financially supported?

6. When was the last time you listened to the material?

7. What did you do for the preservation of the material?

In some cases, the owners actively helped to describe their own archives or the cataloging was carried out by someone who had been active in collecting the recordings.

As far as the naming and the arrangement of the archives are concerned, very often the owners provided some sort of organization to their archives, arranging them in subsections and usually giving each subsection a title related to its content. As a rule, *Gra.fo* followed the organization given to the archive by its owner and arranged and named the archives and their subsections accordingly. When no indication from the owner was given, archives and subsections were named according to some conventions established within *Gra.fo*. Private archives were usually named after the researchers who collected them (e.g. *Archivio "Roberta Beccari", Archivio "Benozzo Gianetti"*), while those belonging to an organization took the name of that organization (e.g. *Archivio "FLOG" – Federazione Lavoratori Officine Galileo –, Archivio "ASMOS" – Archivio Storico del Movimento Operaio e Democratico Senese*). Archives resulting from important geolinguistic enterprises took the name of those enterprises (e.g. *Archivio "Carta dei Dialetti Italiani", Archivio "Atlante Lessicale Toscano"*). The archives' sections (*fondi*) and subsections (*serie*) corresponding to specific research projects were usually named after the topic of the specific research (e.g. *Archivio "Dina Dini", fondo "Emigranti"*), or after the researcher(s) who carried out the investigation (e.g. *Archivio "FLOG", fondo "Andrea Grifoni", serie "Vita di Fabbrica"*). Furthermore, in some cases it was necessary to abandon or emend the arrangement given by the archive's owner in order to overcome idiosyncrasies. For example, a subsection called *"Storie di vita"* ('Life stories') by the archive's owner was renamed *"Storia orale"* ('Oral history'), as it contained interviews on local traditions, peasant and material culture instead of biographies. Another example is offered by the *Archivio "Angela Spinelli"*, a public archive, protected by the Soprintendenza Archivistica and preserved in the Istituto culturale e di documentazione Lazzerini in Prato. It contains the result of an investigation carried out in 1982 by Angela Spinelli and Roger Absalom on the help offered by the rural population of the surroundings of Prato to the British soldiers who escaped from prisoner camps during World War II. In *Gra.fo*, these materials constitute a subsection (*fondo "Lazzeriniana"*) of a larger archive (*Archivio "Angela Spinelli"*) which also includes a preliminary investigation (*fondo "Appendice Lazzeriniana"*) carried out by the same researchers in the same area that has so far constituted a private archive (kept by Angela Spinelli in her house).

In addition, in order to conform to the definition of "open archive" characterizing oral archives produced for scientific research purposes, which are open-ended and reflect the scientific experience of a researcher, *Gra.fo* provides at least one subsection for every archive. In this way, it is possible to accept unexpected subsections without disrupting the arrangement of the archive.[1]

---

[1] For the complete list of the oral archives preserved by *Gra.fo*, please see <http://grafo.sns.it/>.

2.3   Active conservation

Once the audio materials were gathered into the *Gra.fo* laboratory (hosted at the Linguistic Laboratory of Scuola Normale Superiore), the conservation protocol took place. This protocol, inspired by well-established international guidelines (IASA-TC 03) involved i) preparing the original carrier for playback, ii) cleaning and restoring it (if necessary) so as to repair any climatic degradations which may compromise the quality of the signal, iii) choosing an adequate re-recording equipment in order not to introduce further distortions, iv) transferring and archiving the speech signal into the *Gra.fo* database. These activities were carried out differently according to the types of carrier and the recording formats, which often differ in the linguistic, anthropological, and ethno-musicological empirical research. *Gra.fo* mainly dealt with audiotapes, audio reels, Digital Audio Tapes (DAT), Compact Discs (CD), and mass-storage devices. Each of these presented distinct characteristics and thus required a specific approach.

An open-source software system for the preservation and the cataloging of sound archives was developed within the *Gra.fo* project. The software system is called *Audiografo* [5], it combines different technologies and its main components are: *Audiografo PreservationPanel* (*Audiografo PP*), the main function of which is to create preservation copies, and *Audiografo CatalogingPanel* (*Audiografo CP*), for the cataloging of the documents. The active conservation process, supported by the use of *Audiografo PP*, aimed at maintaining all the information represented by the carrier and lead to the creation of two distinct objects: the 'preservation copy' (with high quality uncompressed audio), and the 'access copy' (with lower quality and compressed audio). The former was intended for long-term conservation and contained all the information present in the carrier together with its description and the documentation of the conservation process. The latter, characterized by an equal or inferior sound quality, was further elaborated by the catalogers (the digital signal could be restored – by making use of DSP4 and iZotope software – or manipulated for cataloging purposes, see infra) and finally made available to the end user in the online archive. For a detailed account of the conservation protocols in *Gra.fo* and of the technical equipment employed, see [5] and [6].

3   Cataloging oral documents

Audiografo CP allowed the catalogers to describe both the archives (and their subdivisions) and the single oral documents as follows:

– Information about an archive (or subdivision) – name, place of conservation, existence of the 'Say something about your archive' interview, privacy restrictions, description, motivation of the research, date of joining in the project, owner.
– Information about a single oral document – title, content, date and place of collection, information about the researcher and the speaker, existence of bibliography and accompanying materials, classification of the document, aims of the single recording, keywords.

In the following sections, two key elements of the description of the single oral document will be discussed, namely: the classification of the oral documents and the treatment of the accompanying materials.

3.1 Classification of oral documents

According to the catalographic proposal described in [7], four criteria were used for the classification of oral documents:

1. Typology – The catalogers distinguished between the following options:
   - controlled events (elicited by the researchers and under their direct control, e.g. interviews, answers to a linguistic questionnaire) vs. uncontrolled events (e.g. documents collected with the hidden recording modality, or recordings of folk performing arts events);
   - sung (e.g. lullabies, narrative songs, spontaneous oral poetry) vs. spoken documents (e.g. interviews, narratives, ethno-texts, riddles);
   - formalized (e.g. lullabies, riddles, poems) vs. non-formalized (e.g. interviews, ethno-texts) vs. improvised documents (e.g. narrative songs, spontaneous oral poetry). This distinction implied the analysis of the text format (rhythmic structure, forms of versification, rhymes).

2. Topic – The catalogers could choose among about 130 different topics (such as Agriculture, Anarchism, Animals, Art, Autobiographies, Biographies, Blacksmiths, Carnival, Cinema, Clothing, Coalmen, Cutlers, Dialects and language varieties, Domestic activities, Drug addiction, Emigration, Environment, Exhibitions, Family, Fascism, Fishing, Folk dance, Folk literature, Folk medicine, Folk music and songs, Folk theatre, Folk traditions, Food, Games, Handicraft, Human body, Immigration, Legends, Literature, Local history, Magic, Material culture, Museography, Music festivals, Nazism, Peasant culture, Peasant traditions, Political history, Politics, Postwar period, Pre-industrial society, Prostitution, Racism, Religion, Religious feasts, Rituals, School, Sharecropping, Theatre, Time, Traditional family, Traditional festivals, Traditional food, Traditional jobs, Traditions, Women's condition, Women's history, Work, 1st World War, 2nd World War, etc.). Only one topic per document could be chosen, working as a sort of subtitle stating the main theme of the document. Other relevant (secondary) topics were included in the keyword list.

3. Genre – The catalogers could choose among approximately 40 different genres (such as Answer to linguistic questionnaire, Autobiography, Ethno-text, Image/object description, Interview, Legend, Lullaby, Narrative song, Poem, Political song, Prayer, Proverb, Reading, Recipe, Religious poetry, Riddle, Ritual, Spontaneous conversation, Tale, Theatre, Tongue twister etc.). Creating a fixed taxonomy for such an interdisciplinary project proved to be really difficult, since the available taxonomies were partial (i.e. they referred to a single field of study, such as linguistics, anthropology, oral history, ethnography) and often blurred the boundaries between genres and topics.

4. Language variety – The catalogers could choose among approximately 30 different varieties. According to the taxonomy proposed by Luciano Giannelli [18], [19]), Tuscan varieties were divided into 'urban varieties' (of Florence, Prato, Pistoia, Lucca, Massa, Pisa, Leghorn, Arezzo, Siena, Grosseto), 'areas of influence' (of Florence, Pistoia, Lucca, Pisa, Leghorn, Arezzo, Siena, Grosseto), 'areas of transition' (of Volterra, Massa, Piombino), and other minor varieties (e.g. of the Elba Island). The sociolinguistic motivations for this choice were twofold: a) cities are a vehicle of linguistic identity and usually influence the surrounding areas; b) Tuscany does not have a hegemonic center that can

influence the whole territory of the region (for a detailed account of these issues, see [1] and [8]).

3.2 Accompanying materials

Oral documents need to be carefully interpreted in order to be understood [26], and any relevant note, drawing, diary produced by the researcher before, during and after the data collection constitute a precious resource for correctly interpreting the documents. For this reason, *Gra.fo* devoted great attention to the accompanying materials by digitizing them and making them available to the user together with the sound recordings, the cataloging records, and (if possible) the transcriptions of the documents. The accompanying materials are given in .pdf format and are watermarked, according to a convention established within *Gra.fo*, in order to avoid theft and improper use of the materials.
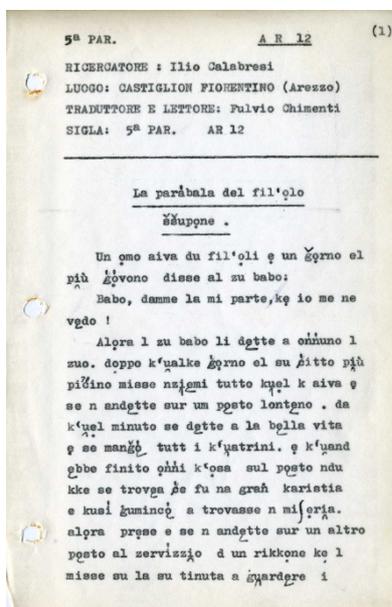
Fig. 1. Example of 'historic' accompanying material from Archivio "Carta dei Dialetti Italiani".

Gra.fo distinguishes between 'historic' and 'interpretative' accompanying materials. The former comprise all annotations and similar documents written by the researcher during the investigation (e.g. the diaries of *Carta dei Dialetti Italiani*). The latter include documents presuming some sort of mediation from the researcher (e.g. orthographic and phonetic transcriptions or unpublished relations). Three documents coming from the accompanying materials of two different archives are given here as examples. The first two come from *Archivio "Carta dei Dialetti Italiani"*, the main linguistic archive preserved in *Gra.fo*. The *Carta dei Dialetti Italiani* archive is a neglected open reels speech archive containing both

answers to phonetic and morphological questionnaires and the oral performance of the Parable of the prodigal son (Lc, 15, 11-32) collected in a significant number of Italian towns by linguists and dialectologists. This important fieldwork, offering a unique database of Italian dialects from the Sixties and Seventies, sank into oblivion because of the death of its founder, Oronzo Parlangèli, and subsequent financial and organizational difficulties. *Gra.fo* succeeded in finding almost all the reels referring to Tuscan fieldwork, together with all the related reports and notes written by the équipe of linguists coordinated by Gabriella Giacomelli [9]. Fig. 1 shows a sheet containing information (date, place, name of researcher, etc.) on a survey carried out in Castiglion Fiorentino (Arezzo) on November 10th 1967 and exemplifies 'historic' accompanying material; Fig. 2 is a phonetic transcription of the Parable of the prodigal son (showing the phonetic alphabet used at that time by Italian dialectologists) exemplifying 'interpretative' accompanying material.

Fig. 2. Example of 'interpretative' accompanying material from Archivio "Carta dei Dialetti Italiani".

The third example comes from *Archivio "Angela Spinelli"*, an oral history archive collected at the beginning of the 80s in Valbisenzio (Prato) by Angela Spinelli and Roger Absalom for the publication of the volume *Il distretto industriale (1943-1993)* of the collection *Prato: storia di una città*. Angela Spinelli moved to a small village in Valbisenzio and interviewed the rural population in order to shed some light on the cultural process that brought about a search for a new socio-political status in the post-war period, and subsequently led to a migration of the rural population towards the city. During and after her investigation, she wrote down information about the informants (among other things, she drew the informants' family trees in order to understand the relations between the different families of

the village) and took notes of the proverbs, the popular religious ceremonies, the objects of material culture, the food, and the illnesses mentioned during the interviews. In the sheet represented in Fig. 3, Angela Spinelli put together all the information gathered about an illness and the relative popular remedies ('interpretative' accompanying material).

Fig. 3. Example of 'interpretative' accompanying material from Archivio "Angela Spinelli", fondo "Lazzeriniana".

4   Transcribing oral documents

The documents that turned out to be interesting from the linguistic point of view (e.g. because they exhaustively exemplified a given variety, or witnessed a disappearing variety) were provided with an orthographic transcription, downloadable as a .pdf file.[2] In *Gra.fo*, orthographic transcription, based on [17] and [20], is intended for representing speech in writing and for being clearly understandable even to non-specialists. Therefore, it implies a compromise between clarity and faithfulness. Furthermore, unlike the systems proposed in [17] and [20], which refer to a single variety, the *Gra.fo* transcription model is applicable to every Tuscan language variety. For this reason, it was decided not to transcribe linguistic phenomena that are constantly adopted by the speaker so as to avoid heavy use of diacritics

---

[2] Thus, the relationship between audio file and transcription appears to be static; the alignment between the audio and the text files is part of our research agenda for the near future.

or special symbols. For example, consonantal weakening phenomena and *raddoppiamento fonosintattico* were not represented for Florentine varieties. Thus, highly variablephenomena presenting sociophonetic value for the given variety were transcribed according to the Italian orthographic conventions. Consequently, every transcription was provided with an introduction mentioning the linguistic phenomena (some of which conveniently transcribed) characterizing the text. In this way, *Gra.fo* fits in the debate on the criteria for the transcription of oral documents carried out by the staff of *Rivista Italiana di Dialettologia* since its establishment. Other conventions were used for transcription: interviewers are indicated by Int.; interviewees are indicated by their initials (e.g. N.S.); participants whose name is unknown are indicated by X.X.; parts omitted for privacy reasons are annotated (e.g. [name], [surname], [job]); parts which are not clearly understandable are substituted by [xxx]; emphasis is marked with *Italics*; dialectal forms are sometimes annotated with the corresponding Italian forms. As for the prosodic domain, the transcription follows the conventions described in [17], [20].

Fig. 4. Example of an orthographic transcription of an interview from Archivio "Dina Dini", fondo "Emigranti".

Fig. 4 shows the first page of the orthographic transcription of an interview from *Archivio "Dina Dini", fondo "Emigranti"*, collected in Pieve Santo Stefano (Arezzo) in 1995-2000, which contains testimonies of informants who migrated to Switzerland, France and Germany in the second half of the 20th century. The lines written in grey correspond to parts of the recording that were censored for privacy reasons. Fig. 5 shows the first page of the orthographic transcription of a narrative song from *Archivio "Roberta Beccari", fondo "Letteratura popolare"*, collected in the area of Leghorn in 1986-1987, which contains testimonies of popular literature, songs and culture.
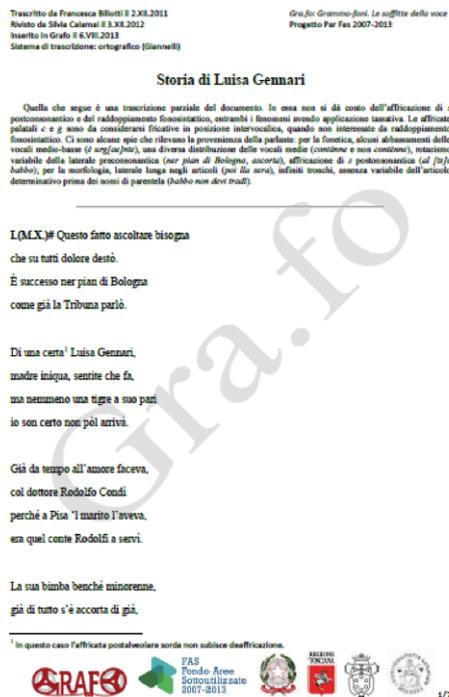
Fig. 5. Example of an orthographic transcription of a narrative song from Archivio "Roberta Beccari", fondo "Letteratura popolare".

5   From the database to the website

The *Gra.fo* MySQL database is made up of 59 interconnected tables, some of which have key constraints. The tables contain information on the fields created for cataloging and for the creation of the preservation copies (which are stored in a specific server archive with a RAID 5 configuration). The digitization and cataloging collaborators interact with the

database through *Audiografo PP* and *AudiografoCP* respectively, which have user-friendly interfaces with drop-down menus, checkboxes and open fields.

The web portal is a navigable interface developed with Liferay, which allows users to query the database and server archive containing the preservation copies, and search all materials collected in *Gra.fo* (cataloging records, .mp3 files, transcriptions and accompanying materials, .pdf files).

The website contains the description of the project, a page devoted to the archives, two pages devoted to the materials' search, as well as the cataloging records of the documents.

The page devoted to the archives contains their names and descriptions, the names of their subsections, and the 'Say something about your archive' interview.

As for the search, users have two options:

  – search by linguistic area (an interactive map allows users to click on the area of interest and access the corresponding records);
  – search by content (users can search by topic, genre and type of document, date and place of the recording, and by language variety).

The cataloging record of each document carries the following information:

  – name and description of the archive (and subsections) to which the document belongs;
  – conditions of access (i.e. whether the document is (partially) restricted for privacy reasons – see infra);
  – title (and alternative title, if present);
  – content;
  – keywords;
  – researcher's name;
  – informant's name, sex, date and place of birth, education level and profession;
  – date, place and setting of the recording;
  – typology;
  – topic;
  – genre;
  – language variety;
  – aim of the recording;
  – bibliography;
  – type of carrier;
  – recording (downloadable in .mp3 format);
  – accompanying materials (downloadable in .pdf format);
  – transcription (downloadable in .pdf format).

All documents concerning the conventions adopted within *Gra.fo* with respect to digitization, restoring, cataloging and transcription protocols are also available online.

The website and the cataloging records are openly accessible but, in order to prevent improper use, only registered users can download .mp3 files, transcriptions and accompanying materials.

Fig. 6 shows a cataloging record from our web portal. The document comes from *Archivio "Roberta Beccari", fondo "Parroci"*, containing interviews with parish priests about popular religiosity collected in the 80s in Maremma (southern part of Tuscany). The screenshot in the top left shows a preview of the document with its title, content, name of the

archive, and the audio recording. By clicking on *Dettagli*, the user can access all the information related to that document (full summary, keywords, topic, genre, linguistic variety etc.) and download the .mp3 file, transcription and accompanying materials (if available), which are shown in the other two screenshots.
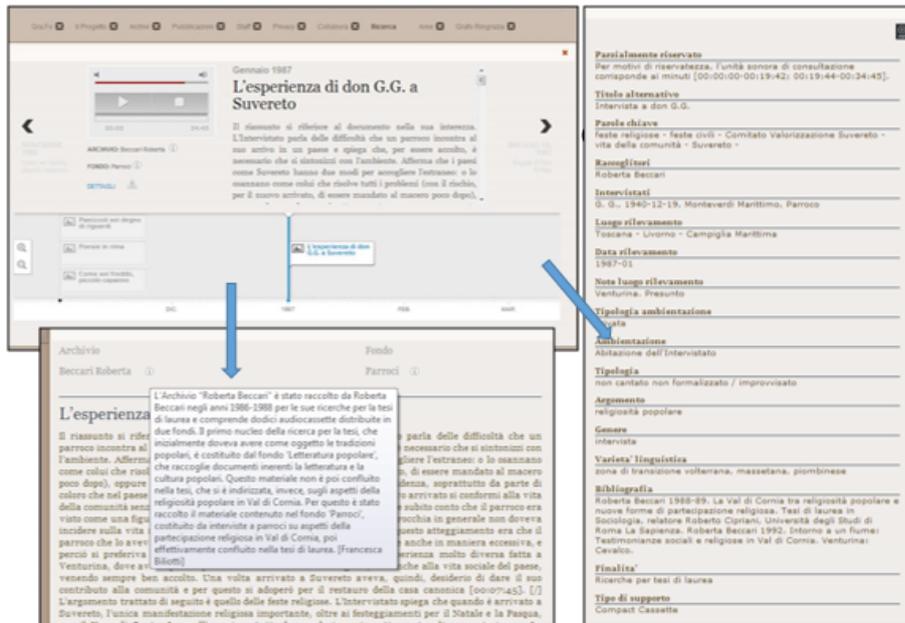

Fig. 6. Example of cataloging record from Archivio "Roberta Beccari", fondo "Parroci".

6   Critical issues

A complex project like *Gra.fo* required the definition of special procedures. Dealing with extremely heterogeneous archives, the *Gra.fo* working group faced a number of critical issues, such as: the relationship between the carrier and the document; the legal treatment of confidential information; the proper treatment of documents containing other documents; the discrepancies between the arrangement given to the archive by its owners and the one adopted within *Gra.fo*. While the latter was dealt with in Section 2.2, the other issues will be discussed in the following sections, together with the solutions adopted.

6.1   From the original carrier to the document

The relation between the original carrier and the document is one of the most problematic issues encountered by the *Gra.fo* team. Occasionally, a document (e.g. an interview, a biography, etc.) occupies various carriers, or portions of carriers (see infra). Yet, it is a single

document and has to be treated as such. Therefore, in *Gra.fo*, the document is considered to be independent from the data carrier(s). In other words, each event, regardless of how long or short it is, and of how many portions of carriers it occupies, corresponds to a document. The carrier is seen as a mere container.

Because the theory of audio conservation was born in the domain of classical music, the literature only refers to (multiple) carriers [22]. In fieldwork, however, researchers used to exploit the carriers fully, thus leaving no portion of tape unrecorded. Consequently, a document can be distributed across various carriers, and one and the same carrier might contain various documents. This depended both on the need to economize and on the fact that (in the past) the transcription and the analysis of the document were valued more than the recording itself.

Within the *Gra.fo* project, it was thus necessary to edit the recordings based on their content. After the creation of the preservation and access copies, which faithfully reproduce the original recording without considering its content, a cataloger edited the access copy and created as many different documents as there were recorded events. The resulting audio file was called 'unit of audio consultation' and was then cataloged, transcribed and made available to the end user on the web portal. Therefore, the preservation and access copies – which are the equivalent of the diplomatic edition – are not accessible to the final users: the object that is offered to public access is the result of an interpretative activity and there is no one-to-one relationship between the two objects.

The process of interpretation and editing was particularly delicate because of the very nature of the *Gra.fo* archive: an interdisciplinary repository composed of very different oral archives whose arrangement was often accessible only to the researcher(s) who collected them. For this reason, within *Gra.fo*, the definition of each basic documental unit was obtained *ex post*, after familiarizing oneself with the archive, the research aims and protocol, the elicitation modalities, and the communicative context of the investigations. Editing thus consisted of a series of conventional, critically informed choices and led to the definition of the following categorization:

1. Interview with questionnaire - the documental unit consists of all the answers given by the same person(s) to the questionnaire in a unitary communicative context.[3]
2. Meeting - the documental unit is the recording of a single meeting.
3. Uncontrolled event, where the researcher is a mere witness and has no influence on what happens (e.g. public performances, or documents collected with hidden recording modality) - the documental unit is the recording of the single event.
4. Elicitation of particular genres (e.g. folk songs, proverbs, riddles, lullabies, etc.) - the documental unit is the single object elicited by the researcher (i.e. the single song, the single proverb, etc.).

---

[3] Identifying a unitary communicative context is not always easy. In the case of geo-linguistic archives, such as *Archivio "Carta dei Dialetti Italiani"* and *Archivio "Atlante Lessicale Toscano"*, ideally, the documental unit should correspond to the length of an entire questionnaire recorded with the same participant(s) in a specific location within a single session. But, the picture appears to be much fuzzier because a) multiple informants of differing age and status were interviewed instead of a single informant, thus introducing significant sociolinguistic variability; b) interviews often ran for more than a day (due to the length of the questionnaire); c) some interviews are "mute" (only attested in transcription); d) some interviews are incomplete; e) there is no uniformity in the elicitation modalities (due to different researchers carrying out the investigation) [11].

For a detailed account of the issue of the relationship between the original carrier and the document, please see [11].

However, sometimes ecdotic issues clash with the materiality of actual data and with the need to reach a compromise between philological coherence and data accessibility. In fact, geo-linguistic archives, such as *Archivio "Carta dei Dialetti Italiani",* are usually made of very long interviews (because the questionnaires used were very long and complex). Therefore, even if compressed in lossy .mp3 files, such interviews pose practical, technical problems to accessibility (end users' ability to download them). For this reason, the interviews of *Archivio "Carta dei Dialetti Italiani",* which occasionally last more than four hours, were divided into smaller parts corresponding to the different sections of the questionnaire (i.e. general information, phonetics, morphology, syntax, lexis, Parable of the prodigal son).

6.2   How Gra.fo dealt with confidential information

Another major problem faced within the *Gra.fo* project is the treatment of confidential information [13]. Many archives included in *Gra.fo* were recorded before the national law on privacy (D. Lgs. 30 giugno 2003 n.196) was passed, so that the informants were not asked to give their authorization for the dissemination of the recordings. As a consequence, the *Gra.fo* team decided to only provide the initials, rather than the full names of the informants.

As far as the content of the recordings is concerned, documents are grouped into three categories:

  – Fully available documents, which do not contain any confidential information and are, therefore, fully accessible on the web portal.
  – Confidential documents, that is, those in which more than 90% of the recording time consists of confidential information); these are accessible on the web portal only through an edited summary. By contrast, the .mp3 file and the accompanying materials are only available for direct consultation in the *Gra.fo* Laboratory.
  – Partially confidential documents, i.e. those containing some confidential data (less than 90%), which are edited in two different versions: a full version, only available for consultation in the *Gra.fo* laboratory, and a partial version (with edited summary), available on the web portal.

When containing confidential data, accompanying materials are made available for consultation in the *Gra.fo* laboratory; transcriptions, on the other hand, are accessible via the web portal once any confidential data has been removed.

For a detailed account of the ethical and legal issues related to use and re-use of research data and to online dissemination of research data, please see [10].

6.3   Documents containing other documents

In some cases, a document that constitutes a single archival unit contains other documents. For example, during an interview on rural traditions, the interviewee might start singing the songs that people used to sing during threshing, thus producing documents that differ substantially from the main document to which they belong. Such circumstances obviously pose serious editing and cataloging problems. On the one hand, one can focus on the main

document, treat it as a single entity (i.e. edit and catalog it as a single cataloging unit) and give all the information about the different documents in its description. On the other hand, one can edit and catalog the main document as a whole and, at the same time, separately edit and catalog every other document it contains. The first choice would yield an inaccurate description of the documents, while the second would lead to the creation of audio duplicates. *Gra.fo* opted for a compromise between these two options: for each document – the main document and those contained within – a separate cataloging unit was created, but only the main document was edited as a unit of audio consultation. In this way, more than one cataloging unit can refer to one and the same unit of audio consultation, thus excluding the risk of overloading the system. The connection between the different documents is explicit in that the main document bears reference to any document contained within it and vice versa. Fig. 7 shows the cataloging record of an interview from *Archivio "Roberta Beccari", fondo Letteratura popolare*. The interview contains three poems and their titles appear in the keyword list (*parole chiave*).

Unfortunately, this procedure had to be suspended due to time constraints. Nevertheless, after this decision was taken, the presence of documents contained in other documents was still marked in an aggregate file and their titles recorded in the keyword list of the main document, so that, in the future, these can cataloged according to the protocol established by *Gra.fo*.

Fig. 7. An example of a document containing other documents from Archivio "Roberta Beccari", fondo "Letteratura popolare".

7 Conclusion and future perspectives

The creation of an archive incorporating the main oral archives of the Tuscan region involved different, interconnected stages of work. It was necessary to lay the foundations for an inter-disciplinary dialog between linguistics, anthropology, computer and archival sciences. Such effort involved eleven people as staff, produced more than 2800 hours of digitized recording and more than 2200 cataloged oral documents containing the voices of little fewer than 300 interviewees and 143 interviewers. While digitization has been completed, the cataloging of the digitized documents is still in progress and will continue through 2017. Yet, the thirty archives preserved by *Gra.fo* are only a small part of the existing Tuscan oral heritage, and further efforts would be necessary in order to ensure the long-term preservation and the dissemination of this valuable cultural heritage.

The *Gra.fo* archive is an important resource for various fields of study, as it provides first-hand data that can be exploited in numerous different ways. One particularly promising way of exploiting oral testimonies is proposed in [24] and [25] and consists in the use of oral documents as intangible cultural assets for the augmentation of a tangible cultural site. Such a novel approach, based on the "Augmented Cultural Heritage" technological paradigm, offers a framework for a "sound tourism" in which the perception of sites is directly transmitted by the voice of the local communities through the creation of an Application model for the fruition of landscape, places, and locations by means of oral archives. Thanks to a dedicated Application, the visitor's mobile device becomes a virtual narrator, recounting stories and anecdotes that can enrich the visit to a site.

8 Acknowledgments

References

1. Agostiniani, Luciano and Luciano Giannelli. "Considerazioni per Un Analisi Del Parlato Toscano." In *L'italiano Regionale*, Michele A. Cortelazzo and Alberto M. Mioni eds., 219-37. Roma: Bulzoni, 1990.

2. Andreini, Alessandro and Pietro Clemente, eds. *I Custodi Delle Voci. Archivi Orali in Toscana: Primo Censimento - Pubblicazioni - Regione Toscana*. Firenze: Regione Toscana, 2007. `http://www.regione.toscana.it/-/i-custodi-delle-voci-archivi-orali-in-toscana-primo-censimento`.

3. Barrera, Giulia, Alfredo Martini and Antonella Mulè, eds. *Fonti Orali. Censimento Degli Istituti Di Conservazione*. Roma: Ministero per i Beni Culturali e Ambientali, Ufficio Centrale per i Beni Archivistici, 1993. `https://www.libreriauniversitaria.it/fonti-orali-censimento-istituti-conservazione/libro/9788871250663`.

4. Benedetti, Amedeo. *Gli Archivi Sonori: Fonoteche, Nastroteche E Biblioteche Musicali in Italia*. Istituzioni Culturali Italiane. Genova: Erga edizioni, 2002. `https//catalog.hathitrust.org/Record/004249013`.

5. Bressan, Federica, Pier Marco Bertinetto, Chiara Bertini, Cristina Bertoncin, Francesca Biliotti, Silvia Calamai, Sergio Canazza and Nadia Nocchi. "Un ambiente informatico per il controllo dei processi relativi alla conservazione attiva in un archivio digitale di corpora vocali." In *La voce nelle applicazioni*, Mauro Falcone and Andrea Paoloni eds., 199-214. Roma: Bulzoni, 2012. `https://www.research.unipd.it/handle/11577/2534352#.WX8q1tOGNE4`.

6. Bressan, Federica, and Sergio Canazza. "A Systemic Approach to the Preservation of Audio Documents: Methodology and Software Tools." *Journal of Electrical and Computer Engineering*, (2013). DOI:10.1155/2013/489515.

7. Calamai, Silvia. "Ordinare archivi sonori: il progetto Gra.fo." *Rivista Italiana di Dialettologia* 35 (2011): 135–64.

8. Calamai, Silvia. "Toscana." *Enciclopedia dell'Italiano*. Roma: Istituto dell'Enciclopedia Italiana, 2011. `http://www.treccani.it//enciclopedia/dialetti-toscani_(Enciclopedia-dell'Italiano)`.

9. Calamai, Silvia and Pier Marco Bertinetto. "Per Il Recupero Della Carta Dei Dialetti Italiani." In *Coesistenze Linguistiche nell'Italia Pre- E Postunitaria*, Tullio Telmon, Gianmario Raimondi, and Luisa Revelli eds., 335–56. Roma: Bulzoni, 2012.

10. Calamai, Silvia, Pier Marco Bertinetto, Chiara Bertini, Francesca Biliotti, Irene Ricci and Gianfranco Scuotri. "Architecture, Methods and Purpose of the Gra.fo Sound Archive." A.C. Addison, G. Guidi, L. De Luca, and S. Pescarin eds., 2:439. Institute of Electrical and Electronics Engineers Inc., 2013. DOI:10.1109/DigitalHeritage.2013.6744801.

11. Calamai, Silvia, Pier Marco Bertinetto, Chiara Bertini, Irene Ricci, Francesca Biliotti and Gianfranco Scuotri. "Building an Open Sound Archive: The Case of the Gra.fo Project." IASA-BAAC Conference "Open doors: new ideas, new technologies". Vilnius, 2013.

12. Calamai, Silvia, Francesca Biliotti and Pier Marco Bertinetto. "Fuzzy Archives. What Kind of an Object Is the Documental Unit of Oral Archives?" In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, 777–85. Lecture Notes in Computer Science. Springer, 2014. DOI:10.1007/978-3-319-13695-0_80.

13. Calamai, Silvia, Francesca Biliotti, Luca Pesini and Pier Marco Bertinetto. "Building an Open Sound Archive: The Case of the Gra.fo Project." In *Proceedings of the 6th international congress on "Science and technology for the safeguard of cultural heritage in the Mediterranean Basin*, 3:264–69. Rome, 2014.

14. Calamai, Silvia et al. "Digital Audio Archives Accessibility." In *Cultural Heritage in a Changing World*, Silvia Calamai, Veronique Ginouvès, and Pier Marco Bertinetto eds.. Springer, In press.

15. Canazza, Sergio, Silvia Calamai, Pier Marco Bertinetto and Amedeo De Dominicis. "A Protocol for the Preservation of Speech Documents Archives: Towards the Digital Curation of the Carta Dei Dialetti Italiani." 4:84–93. Valmar, 2011. `https://usiena-air.unisi.it/handle/11365/41004#.WX8xKNOGNE4`.

16. Codice in materia di protezione dei dati personali [Testo consolidato vigente], Pub. L. No. 196 (2003). `http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1311248`.

17. "Convention for the Safeguarding of the Intangible Cultural Heritage - Intangible Heritage - Culture Sector - UNESCO." UNESCO, 2003. `http://unesdoc.unesco.org/images/0013/001325/132540e.pdf`.

18. Giannelli, Luciano. "Introduzione alla lettura. Il testo come documento di lingua: problemi di rappresentazione e appunti di lavoro." In *Io so' nata a Santa Lucia: il racconto autobiografico di una donna toscana tra mondo contadino e società d'oggi*, Valeria Di Piazza and Dina Mugnaini eds., 43–62. Castelfiorentino: Società storica della Valdelsa, 1988.

19. Giannelli, Luciano. "Italienisch: Areallinguistik VI Toskana." In *Lexikon der romanistischen Linguistik: (LRL) IV*, Günter Holtus and Michael Metzeltin eds., 594–606. Tübingen: Niemeyer, 1988.

20. Giannelli, Luciano. *Toscana*. Pisa: Pacini Editore, 2000. `http://www.pacinieditore.it/toscana/`.

21. Giannelli, Luciano and Valeria Di Piazza. "L'orale Scritto. Una Proposta Metodologica per L'edizione Dei Documenti Orali Del Fondo Roberto Ferretti." In *Fiabe, Leggende, Storie Di Paura... La Narrativa Orale Nel Fondo Roberto Ferretti*, 2:51-71. Archivio delle Tradizioni Popolari della Maremma grossetana. Grosseto, 1995.

22. Mulè, Antonella. "Le Fonti Orali in Archivio. Un Approccio Archivistico Alle Fonti Orali." In *Archivi per La Storia*, 16:111–29. Firenze: Le Monnier, 2003.

23. Petrucci, Livio. "Il problema delle origini e i più antichi testi italiani." In *Storia della lingua italiana*, Luca Serianni and Pietro Trifone eds., 3:5-73. Torino: Einaudi, 1993. `http://www.academia.edu/29725520/Il_problema_delle_Origini_e_i_piu_antichi_testi_italiani.pdf`.

24. Pozzebon, Alessandro and Silvia Calamai. "Smart Devices for Intangible Cultural Heritage Fruition." 333–36. Granada: IEEE, 2015. DOI:10.1109/DigitalHeritage.2015.7413895.

25. Pozzebon, Alessandro, Francesca Biliotti and Silvia Calamai. "Places Speaking with Their Own Voices. A Case Study from the Gra.fo Archives." In *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, 232–39. Lecture Notes in Computer Science. Springer, 2016. DOI:10.1007/978-3-319-48974-2_26.

26. "Safeguarding the Audio Heritage: Ethics, Principles and Preservation Strategy | International Association of Sound and Audiovisual Archives." International Association of Sound and Audiovisual Archives, 2005. `https://www.iasa-web.org/de/safeguarding-audio-heritage-ethics-principles-and-preservation-strategy`.

27. Simonetti, V. "Il Censimento Nell'Analisi Archivistica E Alcune Considerazioni Sulle Fonti Orali." In *I Custodi Delle Voci. Archivi Orali in Toscana: Primo Censimento - Pubblicazioni - Regione Toscana*, Alessandro Andreini and Pietro Clemente eds., 271–81. Firenze: Regione Toscana, 2007. `http://www.regione.toscana.it/-/i-custodi-delle-voci-archivi-orali-in-toscana-primo-censimento`.

Last consultation of URLs: 31/07/2017