Not quite your usual kind of resource.
Gra.fo and the documentation of Oral Archives in CLARIN

Silvia Calamai
Università degli Studi di Siena
Italy
silvia.calamai@unisi.it

Francesca Frontini
Université Paul-Valéry Montpellier 3
France
francesca.frontini@univ-montp3.fr

Abstract

We present some reflections on the documentation of Oral History archives within CLARIN with a focus on their accessibility through the CLARIN Virtual Language Observatory. The case study is constituted by the *Grammo.foni Le soffitte della voce* project, a collection of digitized and catalogued oral Tuscan archives.

1 Introduction

The[1] term "Oral History" generally has a twofold meaning. On the one hand it can refer to a distinct field of social research (i.e., a production process of historical knowledge starting from oral sources); on the other hand it is also used to refer to a methodological tool that is used by different scholars in separate fields (from dialectology to ethnography, from sociology to anthropology). However, despite the existence of these two different senses of the term, there is a general agreement on what the data of Oral History are, namely "spoken accounts on personal history in an interview setting" (Scagliola and de Jong, 2014).

Oral History (taken in both senses of the term) is one of the most data intensive fields of research in the humanities. The availability of a new type of technology (voice recording) and, consequently, a new type of data (audio interviews) made it possible for many historians, linguists and sociologists to shift their interest from written to oral sources (Scagliola and de Jong, 2014). There are numerous separate groups working on Oral History in different countries, some of them operating outside academia. It is not infrequently the case that an individual research group will create its own particular set of metadata for the description of its archives, thus making comparison and interoperability among different resources problematic. For example, whereas French groups generally tend to rely on shared best practices for the description and analysis of Intangible Cultural Heritage resources, which have been conceived precisely for unreleased sound archives (Marcadé et al., 2014), Italian groups still lack a common framework for data description, although some efforts have been made in this direction (Mulè, 1999, 2003; Calamai, 2012).

Oral History archives have very specific documentation issues especially when compared with other language resources such as written corpora. The fact is that unlike other kinds of materials, open reels and compact cassettes are often devoid of documentary support and the original motivations behind their collection may be clear only to the researcher(s) who collected them. Moreover, accessibility also raises specific technical, legal and ethical problems. Research infrastructures such as CLARIN may in part offer solutions for the long term preservation, documentation, visibility and availability of such archives, but ad hoc solutions also have to be considered in order to take into account the specificity of such data.

In this paper we present the case study of *Grammo-foni Le soffitte della voce* (*Gra.fo*), a collection of digitized and catalogued oral archives arising from the Tuscany area. The *Gra.fo* archive with its rich set of metadata is in the process of being documented and entered into the ILC4CLARIN[2]

[2] ILC4CLARIN adopted the LINDAT-DSPACE repository; currently the default LINDAT profile is used, but ad hoc profiles might be implemented.

repository, as well as of being made searchable via harvesting from the CLARIN Virtual Language Observatory. In order to do this, some crucial issues had to be addresses first.

2 CLARIN and Oral History data

CLARIN repositories often have very different documentation practices and use different metadata profiles. Indeed it is through the use of shared concepts[3] that CLARIN is able to gather all records into a meta-catalogue, the Virtual Language Observatory (VLO)[4], and to map the information in each metadata record to the facets of the VLO search interface. In what follows, we shall discuss some of the main issues concerning granularity, searchability, and consistency of the metadata descriptors.

Granularity is a crucial problem in documenting data for CLARIN, as already recognized by Odijk (2014). The documentation practices of many communities – among which Oral History practitioners – foresee the use of a relatively small set of metadata to describe each item in a dataset. For instance, interviews may require the description of all participants in an audio recording, their role and social situation. Moreover, they may provide very specific information concerning the type and the technical aspects of the audio signal. Metadata may follow the hierarchical structure of the data they describe, so that some metadata pertain to a single recording, others to a set of interviews, others again to a whole collection, or even to a whole archive.

The Component MetaData Infrastructure (CMDI) allows metadata curators[5] to describe any information they may find relevant, and also to create vertical and horizontal relations between objects (a text and the corpus to which it belongs, a file and its transcription, a corpus and its browsing interface). Having said this, metadata curators should also take into account the fact that only a limited amount of information should be mapped onto the restricted set of search facets of their local repository, in the first place, and of the VLO eventually. This means that some of the structured information that is stored in the CMDI metadata of a collection may **not be fully searchable** by the users although it remains visible in the full CMDI record.

Oral History data is of great interest to many of the disciplines in the Social Sciences and Humanities, it is therefore important for existing oral archives to be well documented in a CLARIN compatible format and made searchable in the VLO. CLARIN centers already contain a large amount of oral data, including interviews and life stories, and an Oral History profile has been defined in the CLARIN Component Registry. The profile was created ad hoc to describe INTER-VIEW (van den Heuvel et al., 2012; van den Heuvel et al., 2014), a collection of various sets of interviews with soldiers and veterans of the Dutch army. The authors themselves were unsure whether this profile could be generalised to cover other datasets. The full content of the profile is constituted of several nested components and can be consulted in the registry: it contains information about the interview itself (language, location, creators, modality), its content, the interview method followed, the participants, the interviewer and interviewee in particular, the audio recording with its technical characteristics and the annotation file with its features. Notwithstanding this, searches in the VLO show that many profiles and approaches have been used for Oral data **without a consistent approach**, thus making the search of such datasets less transparent for users. More specifically, the aforementioned INTER-VIEW dataset is retrievable as a single entry[6] while in other cases, such as the DoBeS archive[7], the entire hierarchical structure is represented in the VLO, with an entry for each level[8]. The VLO offers a graphical navigation aid in the hierarchy of the data, but it is still not easy for users to orient themselves, especially when some of the hierarchical levels are not well documented.

[3] Shared concepts are stored in the CLARIN concept registry (Schuurman et al., 2016) <https://catalog.clarin.eu/ds/ComponentRegistry/>

[4] < https://vlo.clarin.eu>

[5] By metadata curator we mean here anyone documenting data within a CLARIN registry.

[6] <https://vlo.clarin.eu/record?q=living+oral+history+workbench&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-1163_64_format_61_cmdi>

[7] DoBeS, Documentation of Endangered Languages, contains language documentation data from a large variety of endangered languages from around the world.

[8] <https://vlo.clarin.eu/record?q=DoBeS&docId=hdl_58_1839_47_00-0000-0000-0001-305B-C_64_format_61_cmdi>

3 The Gra.fo dataset in CLARIN: a case study

The Gra.fo project is not entirely devoted to Oral History since its general aim was safeguarding analogue Tuscan sound archives from deterioration and oblivion, independently of any specific research objectives (Calamai et al., 2013). Archives stemming from different sources were digitized and made accessible via an on-line portal[9]. With respect to metadata, the high variety of oral documents inside the repository (from anthropology to linguistics, from ethnography to Oral History to ethnomusicology) encouraged the researches to adopt a very detailed taxonomy which resembled in some ways the CMDI *OralHistory* profile. The current OralHistory profile requires only small adjustments to accomodate the Gra.fo metadata,such as the addition of titles and keywords for interviews.

However other aspects of the metadata sets merit further attention. With regard to content related issues, the most striking metadata issue concerns the field 'language'. In CLARIN it generally refers to a list of official languages with their ISO code, while in Gra.fo a list of vernacular and local languages (dialects) are proposed instead, in order to capture the complexity of the Tuscan sociolinguistic *repertoire* (Maiden 1995).

Another important issue relates to the 'architecture' of the collection. Gra.fo is a collection of oral archives, loaned by various researchers and institutions and divided into sections (It. *fondo*) and subsections (It. *serie*), according to the traditional archival approach (Calamai, Bertinetto, 2014)[10]. An oral archive may represent the entire research career of a single researcher or it may correspond to a research project or may even be associated with an organization which financed the fieldwork. Such intricacies can be mirrored in the name of the Archive itself. Private archives are usually named after the researchers who collected them (e.g., Archivio "Roberta Beccari", Archivio "Benozzo Gianetti"), while those belonging to an organization take 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 take the name of those enterprises (e.g., Archivio "Carta dei Dialetti Italiani", Archivio "Atlante Lessicale Toscano"). Archives' subsections (*fondi*) and their additional subsections (*serie*) correspond to specific research projects and are 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"). Gra.fo itself may be seen as both a dataset and a web interface allowing for the browsing of such datasets. The question of which levels of this hierarchy should be preserved and described in CLARIN is a crucial one, as different metadata sets apply at different levels of the hierarchy.

The level of detail to be adopted is connected to issues relating to licenses, copyright, and attribution. In Gra.fo archives several sections and sub-sections may have been collected by different and various fieldworkers and researchers, who later loaned or even donated their tapes to the Gra.fo initiative. On the other hand, the Gra.fo research group is responsible for the digitization, transcription and curation of such data. Different sub-sections of the Gra.fo archives may have different levels of availability (e.g., full or partial online access, no online access: see Calamai et al. 2016) depending on the interviewees (privacy issues), the wish of the interviewer, and the content (possible copyright issues). Obviously the researcher who collected a given interview should be acknowledged when citing the dataset in a scientific context. At the same time the Gra.fo curators invested a large amount of time in the digitization and documentation of the data, and this effort also deserves recognition. From the practical point of view, a large majority of the original researchers are now retired and in some cases deceased, and thus either the Gra.fo research group or the various associations will act as contact persons. The Gra.fo research group also holds high quality copies of each document (the so called 'preservation copy') with their metadata (Bressan and Canazza, 2013) – neither of which is accessible via web. Finally, the dynamic nature of the dataset should also be taken into account. According to CLARIN practices, datasets should be stable; when major changes occur, a new entry with a different persistent identifier should be created. This facilitates the citability of data and replicability of research results. Yet within Gra.fo, many archives may be in the future enriched with

[9] <https://grafo.sns.it>

[10] See table at <http://grafo.sns.it/archivi;jsessionid=A01D93402E560B107339D38860FDC314>

new materials, and the search interface when accessed at different times may show new results and data.

Given all these considerations, several options are available to describe the Gra.fo data within ILC4CLARIN and the VLO, ranging from maximum to minumum granularity, with intermediate options in between. In the maximum granularity option, all levels of granularity are described as separate entries, namely Gra.fo itself, the various archives, their sections and subsections, and finally the individual entries that are constituted by what is called the documental unit. This is composed by the metadata entry and optionally by the audio recording and additional documentation. A different profile may possibly have to be chosen for each level, this requires an adaptation of the existing OralHistory profile. The Gra.fo web interface would be recorded in CMDI as the browsing interface for each entry and the curators of Gra.fo as the contact persons. In the minimum granularity option, only the Gra.fo interface is described as a single entry; users will find it in the CLARIN repository and will use the Gra.fo search interface to find individual items.

Advantages and disadvantages exist in both cases. In the maximum granularity hypothesis, a very large number of items are inserted in the repository, and the relationships between each dataset and its data-items may become unclear. In the minimum granularity hypothesis, the metadata describing the resource will end up being very generic, the specificity of each archive, section and item being blurred. This might also impact the visibility of the resource, since users of the ILC4CLARIN archive or of the VLO might be looking for a particular topic or data that is present in Gra.fo but cannot be documented in the Gra.fo entry as it does not hold for the totality of the dataset. A typical example may be the language variety since Gra.fo contains data in several local Tuscan vernaculars: in the minimum granularity setting, phonetic or sociolinguistic research on Tuscan varieties may not benefit from the Gra.fo archives being in the VLO, since they would be irretrievable with queries targeting datasets in a specific vernacular variety. Given that the high granularity of language classification in a regional project cannot be reproduced on a large size scale, the chosen profile should at least allow for all the samples of speech which do not conform to the standard variety of a language to be labelled as 'vernacular' or 'non standard'. Moreover, the different varieties used throughout an interview may also have some relationship with the subjects involved in the communicative event (e.g., the interviewer may (try to) use the same linguistic variety of his/her interviewee in order to shorten distances). In this respect, it could be very useful to adopt the solution envisaged in (Marcadé et al. 2014: 30), where some comments on the language used are possible. Similar considerations hold for the topics, which are clearly documented for the individual entries but would become quite generic for the whole resource given the heterogeneity of the resource itself.

4 Conclusion

The documentation of Gra.fo within CLARIN constitutes an interesting case study to test the applicability of current practices and instruments to Oral History and more generally to oral archives data. The problems arising from this investigation which are still to be solved relate to aspects such as categorization (corpus/tool), granularity (collection/item), attribution and citation of intellectual contribution, keeping track of versions and availability.

It became clear that in choosing the right degree of granularity and the right profile for such data, curators should also be guided by the visibility that they want to achieve for their data, both at the level of the local repository and of the VLO. In particular, regarding the latter, it is crucial that a good and meaningful mapping is ensured for the currently available facets so that interested users may be able to find Gra.fo data with ease. In this sense, the work done by (Eckart et al., 2015) to study search behaviors of the VLO users may provide useful guidance to curators.

A solution for the Gra.fo dataset is currently under development, but we believe that common practices should be defined for Oral History and oral archives within CLARIN - considering the fact that a relevant segment of the Oral History community may be interested in depositing their data in CLARIN repositories - so as to allow for a consistent search and navigation of such resources within the VLO.

Reference

[Bressan and Canazza 2013] Bressan, F. and Canazza, S. 2013. A Systemic Approach to the Preservation of Audio Documents: Methodology and Software Tools. In *Journal of Electrical and Computer Engineering* (Article ID 489515): 21.

[Calamai and Bertinetto 2014] Calamai, S. and Bertinetto, P. M. 2014. *Le Soffitte Della Voce. Il Progetto Grammo-Foni*. Manziana: Vecchiarelli.

[Calamai et al 2013] Calamai, S., Bertinetto, P. M., Bertini, C., Biliotti, F., Ricci, I. and Scuotri, G. 2013. Architecture, methods and purpose of the Gra.fo sound archive. *Digital Heritage International Congress (DigitalHeritage), 2013*, vol. 2. pp. 439–439.

[Calamai 2011] Calamai, S. 2011. Ordinare archivi sonori: il progetto Gra.fo, *Rivista Italiana di Dialettologia*, 35, 2011: 135-164.

[Calamai et al 2014] Calamai, S., Biliotti, F. and Bertinetto, P. M. 2014. Fuzzy Archives. What Kind of an Object Is the Documental Unit of Oral Archives? In Ioannides, M., Magnenat-Thalmann, N., Fink, E., Žarnić, R., Yen, A.-Y. and Quak, E. (eds), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection: 5th International Conference, EuroMed 2014, Limassol, Cyprus, November 3-8, 2014. Proceedings*. Cham: Springer International Publishing, pp. 777–85.

[Calamai et al 2016] Calamai, S., Ginouvès, V. and Bertinetto, P. M. 2016. Sound Archives Accessibility. In Borowiecki, J. K., Forbes, N. and Fresa, A. (eds), *Cultural Heritage in a Changing World*. Cham: Springer International Publishing, pp. 37–54.

[Eckart et al 2015] Eckart, T., Hellwig, A. and Goosen, T. 2015. Influence of Interface Design on User Behaviour in the VLO. *CLARIN Annual Conference 2015 in Wroclaw, Poland.*

[van den Heuvel el al 2014] van den Heuvel, H., Oostdijk, N., Sanders, E. and de Lint, V. 2014. Data curations by the Dutch Data Curation Service: Overview and future perspective. *Selected Papers from the CLARIN 2014 Conference*.

[van den Heuvel el al 2012] van den Heuvel, H., Sanders, E., Rutten, R., Scagliola, S. and Witkamp, P. 2012. An Oral History Annotation Tool for INTER-VIEWs. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: ELRA.

[Marcadé 2014] Marcadé, C., Guinard, B., Coulais, S., Davy, Y., Desgrugillers, E., Gasnault, F., Ginouvès, V., et al. 2014. *Patrimoine Culturel Immatériel. Traitement Documentaire Des Archives Sonores in édites. Guide Des Bonnes Pratiques. Guide de Bonne Pratique Pour L'analyse Des Documents Sonores in édits.* Guide des bonnes pratiques - 2014 - Éditions FAMDT

[Maiden 1995] Maiden, M. 1995. *A Linguistic History of Italian*. London: Longman.

[Mulè 1999] Mulè, A. 1999 (ed.) Proposte per la descrizione delle fonti orali. In *Archivi sonori, Atti dei seminari di Vercelli (22 gennaio 1993), Bologna (22-23 settembre 1994), Milano (7 marzo 1995),* Roma 1999, Pubblicazioni degli Archivi di Stato, Saggi 53: 273-292.

[Mulè 2003] Mulè, A. 2003 Le fonti orali in archivio. Un approccio archivistico alle fonti orali, *Archivi per la storia,* 16 (1): 111-129.

[Odijk 2014] Odijk, J. 2014. Discovering Resources in CLARIN: Problems and Suggestions for Solutions Working paper http://dspace.library.uu.nl/handle/1874/303788 (accessed 23 June 2016).

[Scagliola 2014] Scagliola, S. and de Jong, F. 2014. Clio's talkative daughther goes digital: the interplay between technology and oral accounts as historical data. In Bod, R., Maat, J. ter and Weststeijn, T. (eds), *The Modern Humanities*, vol. III. (The Making of the Humanities). Amsterdam, the Netherlands: Amsterdam University Press, pp. 511–26.

[Schuurman et al 2016] Schuurman, I., Windhouwer, M., Ohren, O. and Zeman, D. (2016). CLARIN Concept Registry: The New Semantic Registry. *Selected Papers from the CLARIN Annual Conference 2015, October 14-16, 2015, Wrocław, Poland.* Linköping, Sweden: Linköping University Electronic Press, Linköping universitet, pp. 62–70.

[van Uytvanck et al 2010] van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P. and Gardellini, M. (2010). Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: ELRA.