CLARIN Annual Conference 2021

PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
Proceedings of CLARIN Annual Conference 2021. Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

Programme Committee

Chair:

- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Langa Khumalo, North West University (ZA)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eirikur Rögnvaldsson, University of Iceland (IS)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eirikur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Christophe Parisse, FR
- Thorsten Trippel, DE
- Valeria Quochi, IT
- Zijian Győző Yang, HU
- Efstathia Soroli, FR
- Enikő Héja, HU
- Bence Nyéki, HU
- Angelo Mario Del Grosso, IT
- Olivier Baude, FR
- Kinga Jelencsik-Mátyus, HU

CLARIN 2021 submissions, review process and acceptance

- Call for abstracts: 19 January 2021, 1 March 2021
- Submission deadline: 28 April 2021
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 16-17 June 2021
- Notifications to authors: 22 June 2021
- 35 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event.

Table of Contents

Less is more when FAIR. The minimum level of description in pathological oral and written data

Rosalba Nodari
University of Siena, Italy
rosalba.nodari@unisi.it

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

Henk van den Heuvel
Radboud University, Netherlands
H.vandenHeuvel@let.ru.nl

Abstract

This paper presents a case study under the DELAD initiative, on the basis of two different types of data originating in a former neuropsychiatric hospital in Italy: a collection of oral interviews recorded in 1977 by Anna Maria Bruzzone inside the hospital, and a long diary written by a schizophrenic patient in the Seventies. Given the vulnerability of the subjects involved, and the distance in time from the data collection, not all the audio and written material may be accessible). The aim of this work is to address some of the challenges in archiving and storing legacy data referring to vulnerable people in European infrastructures, and to present a minimum set of metadata that can be accessed for further research, according to the FAIR principles.

1 Introduction

Data collections with written or spoken accounts of people with mental disorders are not easy to obtain for research purposes. Considerable effort (and serendipity) is required to find them, but that is only the first hurdle. Permission must then be obtained to use the records for study. Technical challenges may also arise in order to convert the material into digital format to make the data interoperable for analysis with modern technological means. More complications can emerge if the material needs to be shared with other researchers. At all these stages ethical, technical and General Data Protection Regulation (GDPR) issues need to be dealt with. To help researchers do this, the DELAD initiative (see http://delad.net) was established. DELAD stands for Database Enterprise for Language And speech Disorders, (notably: "delad" is Swedish for "shared"). DELAD is an initiative to share corpora among researchers of the speech of individuals with communication disorders (CSD). This is done in a GDPR-compliant way and at secure repositories in the CLARIN infrastructure. DELAD organises workshops focusing on how such corpora can be made shareable with other researchers (Lee et al, 2021). To this end DELAD cooperates with CLARIN data centres such as The Language Archive at the Max Planck Institute (https://archive.mpi.nl/tla/), and Talkbank at CMU (https://talkbank.org/).

In order to offer a corpus of disordered speech shareable with other researchers, the University of Siena research group decided to join the DELAD initiative. The Arezzo Neuropsychiatric Archive preserves different linguistic materials of people with mental disorders. The archive is of notable interest because it offers written and spoken documents of unmonitored speech, unlike other CSD corpora, in which speakers usually perform particular linguistic tasks. It is, furthermore, rich in documentation, offering multiple types of data, and in its population sample. Nevertheless, the corpus in question poses several problems which must be taken into account. We are facing, in fact, not only ethical issues, but also legal issues, given the historical nature of the corpus. The need to find a balance between the safeguarding of vulnerable subjects and the necessity of offering the scientific community a suitable corpus for future research was the reason for starting a feasibility study for storing and archiving the linguistic material present in the Arezzo neuropsychiatric hospital archive into the Language Archive at the Max

Plank Institute for Psycholinguistics in Nijmegen. In the following sections we first offer a description of the Arezzo neuropsychiatric hospital archive. We then focus on the creation of metadata for the written specimens preserved in the Archive according to the best practices for depositing material in the repository bundle of the Language Archive. Finally, we draw conclusions.

2 The historical archive of the neuropsychiatric hospital in Arezzo: its composition, and why it can be important for speech research

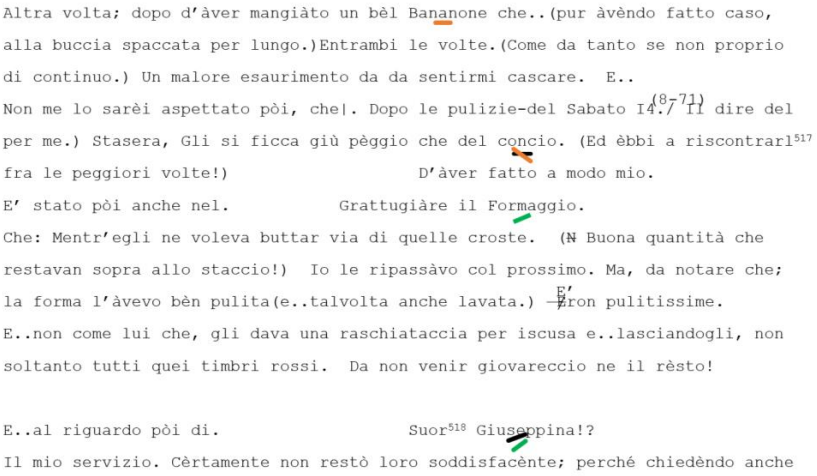
The historical archive of the former Arezzo neuropsychiatric Hospital (Italian acronym ONP) is kept at *Palazzina dell'Orologio* in the Department of Education Sciences, Human Sciences and Intercultural Communication at Siena University, in the Arezzo campus, where the Arezzo sanatorium/mental hospital was originally located. After years of negligence and abandonment following the closure of the Hospital in 1990, the hospital documentation, starting in 1999, was retrieved and catalogued by the University of Siena, in agreement and collaboration with the regional health authority, the owner of the documentation, and the Archival and Bibliographic Superintendence of Tuscany (SabTo). This reorganization made it possible to reconstruct the various original archival series, grouped in two sections and corresponding to the functional sections of the hospital administration: the Directorate for administrative and health affairs, and the Bursar's office for economic management. In 2004, the full inventory of the archive was published (within the series Progetto Archivi of the Province of Arezzo) and can now be consulted online within the "*Carte da legare*" project of the General Directorate for archives (Gherardi, Montani 2004).

The Archive is now open and accessible to university students, researchers and citizens seeking information on their family history. It is composed of about 1500 elements, including files, registers, envelopes, notebooks, and filing cabinets. According to an Agreement renewed in 2019, the University of Siena manages the safekeeping and access in the consultation of the stored material. In addition, a scientific Committee is in charge of enhancing, promoting the study of collections, organizing scientific events and coordinating projects. One of the aims of the Committee is to recover missing archives regarding the neuropsychiatric hospital, and to involve citizens in documentary research and memory conservation, in order to build and preserve hybrid archives (audio, audio-video, paper-based).

Since 2016 several private archives of public figures, who in various ways had a close connection to the Institute of Arezzo, have been traced and collected. Among these is the Bruzzone archive, created by the former teacher and independent researcher Anna Maria Bruzzone. Paola Chiama, Bruzzone's granddaughter and custodian and depositary of Bruzzone's work, decided to progressively donate the researcher's whole archive. Anna Maria Bruzzone conducted various interviews with the patients of the hospital of Arezzo in the summer of 1977. The transcriptions of these interviews were later collated by Bruzzone into the book entitled *'They called us mad. Voices from a mental hospital*' (Bruzzone 1979), republished in 2021. Thanks to her donation, the Arezzo Archive now preserves 19 audiocassettes containing the recordings of the interviews from 34 patients (16 men and 18 women), and 17 other cassettes on different topics. The tapes are associated with the original transcriptions, in different developmental stages – from the Bruzzone handwritten transcription to the published version. The "Anna Maria Bruzzone oral Archive" –which was declared to be of considerable historical and cultural interest in 2018 – is kept at the historical Archive of the former ONP of Arezzo, to which it is ideologically and strongly connected.

Additionally, the archive's nucleus consists in the medical records of patients admitted to the mental and neurological wards. These medical records represent a permanent series in which the files are organized alphabetically and are now preserved in the Directorate section. They consist of two subseries, one devoted to the "mental ward", and the other one to the "neurological ward". Totally, the archive preserves 11,935 medical records of patients hospitalized in the mental unit (see below) and 19,129 medical records regarding the patients in the neurological unit. The medical records are of particular interest in that they contain patients' personal files, such as vital records, administrative and health documents, decrees of partnership and interdiction, medical certificates, family histories, clinical charts, etc., and, until the 1970s, at least one photograph of the patient. In some cases, the charts preserve sections that were directly written by the patients, such as private correspondence, poems, drawings and diaries or autobiographies.

The diary of a former patient is of particular interest. It was meticulously preserved by Fabio Marzi, a psychiatrist who worked at ONP and was loaned to the ONP archive by Gaetano Marzi, the psychiatrist's son. The diary is typewritten, paginated and bound by the author himself, and consists of 323 pages. At the end, it contains a detailed subject index organized according to the places he visited during his life, and to the events he described, in addition to the meetings with people whose stories are told in the diary. The index is organized according to different underlining colours which occur in the text (i.e., the different colours refer to groups of people associated with the Neuropsychiatric Hospital, to family and personal interest, and to different places: see Fig. 1).


Altra volta; dopo d'àver mangiàto un bèl Bananone che..(pur àvèndo fatto caso,
alla buccia spaccata per lungo.)Entrambi le volte.(Come da tanto se non proprio
di continuo.) Un malore esaurimento da da sentirmi cascare. E..
Non me lo sarèi aspettato pòi, che|. Dopo le pulizie-del Sabato 14.7 (8-71) Il dire del
per me.) Stasera, Gli si ficca giù pèggio che del concio. (Ed èbbi a riscontrarl[517]
fra le peggiori volte!) D'àver fatto a modo mio.
E' stato pòi anche nel. Grattugiàre il Formaggio.
Che: Mentr'egli ne voleva buttar via di quelle croste. (N Buona quantità che
restavan sopra allo staccio!) Io le ripassàvo col prossimo. Ma, da notare che;
la forma l'àvevo bèn pulita(e..talvolta anche lavata.) E' ron pulitissime.
E..non come lui che, gli dava una raschiataccia per iscusa e..lasciandogli, non
soltanto tutti quei timbri rossi. Da non venir giovareccio ne il rèsto!

E..al riguardo pòi di. Suor[518] Giuseppina!?
Il mio servizio. Cèrtamente non restò loro soddisfacènte; perché chiedèndo anche

Figure 1. Specimen of the diary (page 59)

Thanks to the medical records, it is now possible to reconstruct personal information about the author. P.A. was a male born in 1916, single, with basic literacy skills and who wanted to join the Church. He was hospitalized two times, the first time from July 1952 to March 1955, the second one in July 1963. Later he returned on his own volition to the hospital, when it was under the direction of Agostino Pirella. On this occasion, he left the hospital only once, in October 1977 for four days.

All this heterogeneous material, by virtue of its nature, needs particular attention in the management, metadata creation and conservation. According to data's peculiar importance (i.e., personal data of vulnerable people) it is indeed crucial to find the right balance between research and the protection of privacy, in order to permit the transmission of knowledge and freedom of research, while maintaining the protection of personal data. In the next section we will propose a minimum set of metadata, in order to make the archive suitable and accessible for researchers interested in disordered speech.

3 The metadata

The Language archive uses the CMDI (CLARIN Metadata Infrastructure) framework as a standard for its descriptive metadata (Broedet et al. 2008; de Vriend et al. 2013). According to TLA deposit manual (https://archive.mpi.nl/tla/deposit-manual-tla), for the Arezzo ONP archive the lat-corpus metadata profile is used as a baseline. The web-based deposit system of TLA includes a webform where the existing metadata profile can be edited for all the relevant collections, sub-corpora and bundles. The CMDI is of profitable use because the CLARIN infrastructure offers researchers the possibility of using ready-made standard component and profile metadata that can be easily adapted to specific linguistic collections. Additionally, the possibility of inserting metadata using the Language Archive web-based interface guarantees a user-friendly tool that does not require any competence in XML syntax. The web interface thus allows to split the work among different collaborators who can then compile the metadata profile online, after having obtained a registered account.

Nevertheless, given the peculiar nature of the Arezzo ONP archive, it is necessary to conduct a preliminary analysis of the architecture of the reference corpus with the aim of selecting the appropriate metadata components and profile from the existing set of metadata. In this respect, at least three crucial

issues have to be taken into account. The first issue is the heterogeneity of the archive. As was mentioned above, the Arezzo Neuropsychiatric Hospital archive contains not only the oral interviews collected by Bruzzone, but also written material (i.e., a private diary). In this regard we have created at the TLA a general collection, called Arezzo Neuropsychiatric Hospital, that, in the future, will contain bundles and other subcollections (such as the Anna Maria Bruzzone archive). In this case, it is then more advisable to consider both granularity – that is, combining components in order to cover just one aspect at a time – and modularity, in order to create a set of metadata that can be suitable for different resources at the same time. For this reason, the first level of metadata profile will thus apply to the collection (i. e. Arezzo ONP archive) and it will contain generic metadata profile, such as Location and Language. Some of these basic components will ~~then~~ be reused for other sublevels of the general collection. Next, different appropriate levels of description will apply to bundles and subcollections.

The second issue to take into account is the peculiar nature of the archive. According to the CMDI best practice guide (https://www.clarin.eu/content/cmdi-best-practices-guide), good component metadata should be "as generic as possible and as specific as needed". This holds particularly true when creating a profile for a corpus that, by virtue of its peculiar nature (i.e. speech and written specimens of vulnerable people), is suitable only for restricted access. Corpora of disordered speech, in fact, usually enclose special categories of personal data (GDPR) such as health information about the patients (see for example van den Heuvel et al. 2020 for a similar case). Thus, metadata creation should not offer sensitive data, such as medical diagnoses, even if these might be useful for other researchers. Nevertheless, the hybrid nature of the archive permits to offer different levels of description depending on specific Bundles and subcollections. For example, the sub-collection Bruzzone archive will contain different Bundles with the complete transcriptions and the different edited versions made by A.M. Bruzzone so that the interviews could be prepared for publishing. However, because the metadata for the Bundle applies to all files within the Bundle, a decision must be made in order to offer a metadata profile that can be applied to all the different versions of the transcription. In fact, the original verbatim transcription contains real names of the patients, that cannot be made publicly available. In this case, the metadata element Actor can be kept anonymous or can contain only the pseudonyms used in the books.

Lastly, the historical nature of the archive makes it necessary to deal with uncertainty. For example, for same patients a medical diagnosis is not available, or the demographic information is incomplete. For this reason, we aim at providing a set of metadata in which it will be possible to infer if some information is not present, rather than omitting possible additional information that could be helpful for other researchers. Following these considerations, we suggest a possible solution for the metadata that will be used for the written material of the archive, that is, at the moment, the schizophrenic patient's diary:

- Name: Diary of a schizophrenic patient
- Title: Diary of a schizophrenic patient / Diario di un paziente schizofrenico
    - Description: Scan and transcription of the personal diary of a schizophrenic patient hospitalised in the Arezzo ONP from July 1952 to March 1955, in July 1963 and, later, under the direction of Agostino Pirella, in the Seventies. The patient wrote the diary his hospitalizations. The diary is typewritten, paginated and bound by the author himself, and consists of 323 pages. At the end, it contains a detailed subject index organized according to the places he visited during his life, and to the events he described, in addition to the meetings with people whose stories are told in the diary. The index is organized according to different underlining colours which occur in the text (i.e., the different colours refer to groups of people associated with the Neuropsychiatric Hospital, to family and personal scope see comment above, and to different places).

- Location
    - Continent: Europe
    - Country: Italy
    - Region: Tuscany
- Project
    - Name: Arezzo ONP archive
    - Contact: […]

    - Description: The corpus contains some of the material preserved in the Arezzo ONP archive. The archive is composed of about 1500 elements, including files, registers, envelopes, notebooks, and filing cabinets, and it documents the history of the Arezzo mental health institution. Along with these materials, the archive comprises different collections of linguistic interest. One of these collections is the Bruzzone archive, created by the former teacher and independent researcher Anna Maria Bruzzone, who conducted various interviews with the patient at the Arezzo ONP. Other collections preserve different specimens that were directly written by the patients, such as private correspondence, poems, drawings and diaries or autobiographies.
- Content
    - Genre: diary
    - SubGenre: spontaneously written
- Content_Languages
    - Content_Language:
    - Id: ISO639-3:ita
    - Name: Italian
- Actors
    - Actor:
    - Role: Patient
    - Name: PA
    - FamilySocialRole: Unspecified
    - EthnicGroup: Unspecified
    - BirthDate: 1916
    - Sex: Male
    - Education: Basic literacy skills
    - Age:
- Resources
    - WrittenResource
        - Date: Unspecified
        - Type: Scan
        - Format: application/zip
        - Size:
    - WrittenResource (Written resource)
        - Date: Trascription
        - Type: Scan
        - Format: application/zip
        - Size:

Following the Deposit Manual of The Language Archive, the metadata will be provided through the online form of the and will be uploaded with the corresponding file, thus creating a bundle with all the resources of the different subcollections.

4 Conclusion

Archiving, managing and sharing corpora of disordered speech is a challenging task, and when managing with legacy data can add even further complications. The option of offering full information when providing metadata will have to, in fact, deal with GDPR issues regarding the protection of special categories of personal data. This feasibility study for storing the linguistic material from the Arezzo ONP archive into the Language Archive aims precisely at striking a good balance between acting in a GDPR-compliant way, while still offering an essential, but still useful, set of metadata suitable for other researchers. However, an effort should be made so as to offer the research community more sensitive information, such as medical diagnoses in the form of documents with restricted access. It is hoped that, at the conclusion of the present work, further studies will be undertaken to assess the feasibility of accessing currently restricted data within the Arezzo ONP archive. DELAD will assist in ensuring compliance to the GDPR guidelines for data protection, transparency, and accessibility.

References

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016: 1-88. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (Accessed 14-04-2021).

Broeder, D., Declerck, T., Hinrichs, E., Piperidis, S., Romary, L., Calzolari, N. and Wittenburg, P. 2008. Foundation of a component-based flexible registry for language resources and technology. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).*

Bruzzone, A. B. 1979. *Ci chiamavano matti. Voci da un ospedale psichiatrico.* Einaudi, Torino.

Bruzzone, A. B. 2021. *Ci* chiamavano matti. Voci da un ospedale psichiatrico. Nuova edizione a cura di Setaro, M. and Calamai, S. Il Saggiatore, Milano.

de Vriend, F., Broeder, D., Depoorter, G., van Eerten, L. and Van Uytvanck, D. 2013. Creating & Testing CLARIN Metadata Components. In *Language Resources & Evaluation (LREC)*, 47: 1315–1326.

Gherardi, S. and Montani, P. 2004. Inventario dell'archivio storico dell'ospedale neuropsichiatrico di Arezzo. Le Balze, Arezzo. http://www.cartedalegare.san.beniculturali.it/fileadmin/redazione/inventari/Arezzo_Ospedale-Neuropscichiatrico.pdf

Kelli, A., Lindén, K., Vider, K., Kamocki, P., Birštonas, R., Calamai, S., Labropoulou, P., Gavriilidou, M. and Straňák, P. (2019). Processing personal data without the consent of the data subject for the development and use of language resources. In *Selected papers from the CLARIN annual conference 2018*, Pisa, 8-10 October 2018. Linköping University Electronic Press: 72-82.

Lee, A., Bessell, N., Van den Heuvel, H., Saalasti, S., Klessa, K., Müller, N., and Ball, M. J. 2021. The latest development of the DELAD project for sharing corpora of disordered speech. Accepted for: *Clinical Linguistics & Phonetics.*

van den Heuvel, H., Kelli, A., Klessa, K., and Salaasti, S. 2020. Corpora of Disordered Speech in the light of the GDPR: two use cases from the DELAD Initiative. In *Proceedings of the 12th Language Resources and Evaluation Conference*: 3317-3321.