Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/101864

An Oral History Annotation Tool for INTER-VIEWs

Henk van den Heuvel (1), Eric Sanders (1), Robin Rutten (1),

Stef Scagliola (2), Paula Witkamp (3)

(1) CLST, Radboud University Nijmegen, The Netherlands
(2) Veterans Institute, centre for knowledge and expertise, Doorn, The Netherlands
(3) DANS, Data Archiving & Networked Services, The Hague, The Netherlands

E-mail: {H.vandenHeuvel,E.Sanders}@let.ru.nl, RobinRutten@student.ru.nl,
scagliola@eshcc.eur.nl, paula.witkamp@dans.knaw.nl

Abstract

We present a web-based tool for retrieving and annotating audio fragments of e.g. interviews. Our collection contains 250 interviews with veterans of Dutch conflicts and military missions. The audio files of the interviews were disclosed using ASR technology focussed at keyword retrieval. Resulting transcripts were stored in a MySQL database together with metadata, summary texts, and keywords, and carefully indexed. Retrieved fragments can be made audible and annotated. Annotations can be kept personal or be shared with other users. The tool and formats comply with CLARIN standards. A demo version of the tool is available at http://wwwlands2.let.kun.nl/spex/annotationtooldemo.

Keywords: Audio mining; Annotation; Crowd sourcing

1.   Introduction

Interview data can be used in a number of ways, such as comparative, follow-up, secondary analysis, replication, methodological advancement, and for teaching and learning. Experiences with the re-use of interview data show that there is an enormous potential for this type of data. The field of Oral history shows many examples of easily accessible web collections in the US[1] and in the UK[2]. Especially in the field of interview data related to the Second World War and other military conflicts multidisciplinary research is carried out. In the Netherlands the collection of an extensive corpus of over 1000 biographical interviews with veterans of various military missions of the Dutch Armed Forces was recently completed[3], half of which is publicly accessible. Other initiatives in the Netherlands which show how access to Oral History collections can be improved by the use of modern language and speech technology are in progress[4]. In a European framework the disclosure of audiovisual collections with the help of advanced search technology (image, text, sound) is addressed in a recent project called AXES[5].

Rather than by just making text and audio files of the data publicly available online, the humanity's research community is helped by tools which allow searching the data on a fragment level and making it possible to add annotations through a user friendly web interface. For this purpose we used our earlier experience with the development of the fragment fitter tool. This was designed to assist researchers in extracting audio fragments from interviews and linking them to their digital texts to realize a so called Enhanced Publication (see Van den Heuvel, et al., 2010). The new tool was equipped with the following features:

- Retrieval of interviews and fragments of interviews based on Automatic Speech Recognition output

- Audio playback for retrieved fragments

- Metadata of all interviews

- Transcription of audio segments

- Annotations to fragments to be added by registered users

- A user administration to restrict the transcription & annotation facilities to registered users

- Adjustment of a fragment's start and end point

- Advanced search options

- The tool is fully compliant with CLARIN-NL standards

The tool has some resemblance to the Integrated Digital Tool described by Sing & Ambati (2010) and to the Ariadne system (Menke & Mehler, 2010), but these are more focused on corpus management and less to data exploration and annotation by researchers.

2.   Background

Over the years, the project team partners carried out a number research projects in which interviews with veterans were used and re-used for qualitative analyses.

---

[1] e.g. http://historymatters.gmu.edu/mse/oral/online.html
[2] e.g. http://www.oralhistory.org.uk/resources.php
[3] http://www.veteraneninstituut.nl/projecten/interviewproject
[4] see e.g. http://www.verteldverleden.org/
[5] http://www.axes-project.eu/

For the first project, the Veteran Tapes VP (VT-VP, was. a fragment fitter tool was developed and used by researchers from different disciplines in six so called enhanced publications. These means that the citations form the interviews which have been chosen by the researchers and have been integrated in their digital text can be linked to audio-fragments By clicking the link the citation becomes available in audio with the associated text (transcription) and meta-data. In 2010 the articles created by the researchers were published in a book and a concomitant Enhanced Publication[6]. In a second project "Living Oral History Workbench" (LOHW), tools were developed to index 250 interviews with relevant search terms using Automatic Speech Recognition, and to annotate retrieved fragments in a Wiki-like environment[7]. In the third project, INTER-VIEWs[8], the 250 interviews and the annotation tool were made compatible to CLARIN-standards focusing on issues as durable storage, persistent identification, search and harvesting mechanisms, annotation tools, and privacy protection mechanisms. The aim of the project was to deliver generic, standardized solutions for the optimal scientific use of interview data. Thus, the project will enable researchers to review and share each other's views on the data. This principle is represented in the title of the project: INTER-VIEWs

The INTER-VIEWs collection contains:
- 120 interviews presenting a range of experiences and frames of reference of Dutch military during the Second World between 1935-1945.
- 100 interviews with veterans of the Dutch East Indies decolonization war 1945-1949 . This collection represents a large diversity of experiences at the local level with guerilla warfare.
- 30 interviews with veterans of the New Guinea conflict with Indonesia. This is a relatively unknown conflict with very interesting elements (soldiers left in uncertainty and isolation, and the pressure of the international community to decolonize the area)

Each interview lasts between 1 and 3 hours. About 20 out of the 250 interviews were manually transcribed.

The interviews are hosted by DANS using persistent identifiers. Audio files are in wav-format. All contain extensive metadata in an electronic form and elaborated textual summaries divided into time blocks of about 10 minutes. The metadata can be accessed through DANS's EASY system[9].
CLARIN-requirements (Odijk, 2010) were met by implementing a number of actions:
- All interviews were persistently stored at DANS

- Persistent Identifiers are used to refer to the individual interviews
- The metadata components of the interviews were registered in the CLARIN's metadata component registry[10] and linked to ISOcat categories[11]
- The same was done for the metadata of the Oral History Annotation Tool (e.g. tool name, owner, makers, description, input, output, availability)

3. Description of the tool

3.1 Technical design

The tool is designed as an asynchronous web application. This means we did not want the user to navigate from page to page , with the consequence of having to wait for complete page reloads. Instead, the application's update is partially determined by the user interactions, while the data can be requested on the fly.

To accomplish this the application has made heavy use of the scripting language Javascript by employing the well-known open source JQuery javascript library. The set of methods we used , for the exchange of data between the frontend (application in the web browser) and the backend of the application (the server), are also referred to as AJAX.

The backend of the application, being responsible for the communication with the database and sending data back to the end user, is in the PHP5 programming language. Since there is much data stored in the MySQL database, the database tables are carefully indexed to retrieve data quickly and efficiently.

The backend is designed from scratch using the Model-View-Controller (MVC) architecture pattern. In short, this means a request of the user is routed to the right controller, that will either use XML or HTML to build up a document using the data retrieved by the models. This document is processed by the frontend to fit in the HTML Document Object Model (DOM).

A screenshot of a retrieval window is shown in Figure 1.

Since access to a number of interviews is restricted to selected researchers, the tool has an administrator component in which the permissions of various users can be set and monitored. The tool comes with an in-built manual. The interface language is English or Dutch and controlled by a language switch.

For demo purposes an open version of the tool was created allowing access to only those interviews for which a consent form allowing open access is available (about 140 out 250).

3.2 Finding relevant Interview fragments

Each interview is divided into fragments. A fragment usually contains one spoken phrase. The Annotation Tool offers the option to search for fragments using keywords, or using interview IDs and time stamps. In the keyword

Fig. 1: Retrieval screen from the Oral History Annotation Tool

Figure 2: Zoom-in on annotation section of the tool

mode advanced search options are available, such as word combinations, word exclusions, word stemming.

The retrieved interview fragments are shown in the left window (see Fig. 1). Upon selecting a fragment, the relevant information is shown in the large window to the right.

3.3 View fragments

The following information per fragment is available:

- **Transcription** The transcription of the selected fragment. This information is not displayed when the transcription is generated by automatic speech recognition.

- **Sound** Here one can listen to the selected fragment.

- **Metadata** More information about the interviewee and the mission.

- **Summary** Summary of the part of the interview (10 minutes) where the selected fragment belongs to.

- **Keywords** Keywords for the part of the interview (10 minutes) the fragment belongs to.

- **Added annotations** An overview of the annotations to this fragment or an overlapping fragment added by the user and others. One can also add new annotations (possibly with one or more attachments).

- **Added transcriptions** An overview of the transcriptions of this fragment or an overlapping fragment added by the user and others. In this window one can also add a new transcription.

- **Official reference** Reference to the interview on the DANS website (via persistent identifier).

3.4 Add Transcriptions or annotations

With the tool, transcriptions and annotations can be added to a selected fragment. The facility also allows to upload files (with maps, pictures, text, in various formats) to substantiate the annotation.

The user may choose if she/he wants to make the annotation publicly available by checking or unchecking the 'Make publicly available' box.

Only public annotations/transcriptions are visible to others.

Fig. 2 shows an excerpt of a screen showing an annotation with an uploaded bitmap file containing a map.

4. Future work

Further planned extensions and modifications to the tool include:

- The option to navigate through the audio of the full interview

- Extend search facility to metadata, annotations, summary texts

- Integrate the tool with the fragment fitter so as to make it suitable for Enhanced Publications

- Visualisation by a timeline to show the chronological order of inserted annotations

- Introduce a shop cart in which a user can collect relevant fragments for his/her own use

- Employ the tool for other audio collections

One of our main efforts presently is to recruit a community of researchers who will use the tool, present us with feedback for further improvement of the tool and stimulate each other to use it in order to increase the number of annotations to a valuable collection of added material.

5. Relevant links to the tool

- Restricted, complete version for researchers:
  http://watveteranenvertellen.nl/annotationtool/?set_lang=english

- Demonstration version for trying out:
  http://wwwlands2.let.kun.nl/spex/annotationtool demo

- Showcase page at CLARIN-NL:
  http://www.clarin.nl/showcase/inter-views/inter-views

6. Acknowledgements

7. References

Odijk, J. 2010. The CLARIN-NL project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC-2010*, pp. 48-53. Malta.

Singh, A.K, Ambati, B.R. (2010). An integrated tool for accessing language resources. In *Proceedings LREC2010,* Malta, pp. 190-195.

Menke, P., Mehler, A. (2010) The Ariadne system: a flexible and extensible framework for the modeling and storage of experimental data in the Humanities. In *Proceedings LREC2010, Malta,* pp. 944-948.

Van den Heuvel, H., Van Horik, R., Scagliola, S., Sanders, E., Witkamp, P. (2010). The VeteranTapes: Research corpus, fragment processing tool, and enhanced publications for the e-Humanities. In: *Proceedings LREC2010, Malta,* pp. 2687-2692.