The VeteranTapes:

Research corpus, fragment processing tool, and enhanced publications for the e-Humanities

Henk van den Heuvel (1), René van Horik (2), Eric Sanders (1), Stef Scagliola (3), Paula Witkamp (2)

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

E-mail: {H.vandenHeuvel | E.Sanders}@let.ru.nl, {Rene.van.Horik | Paula.Witkamp}@dans.knaw.nl, S.Scagliola@veteraneninstituut.nl

Abstract

Enhanced Publications are a new way to publish scientific and other results in an electronic article. The advantage of EPs is that the relation between the article and the underlying data facilitate the peer review process and other quality assessment activities. Due to the link between de publication and the research data the publication can be much richer than a paper edition permits. We present an example of EPs in which links are made to interview fragments that include transcripts, audio segments, annotations and metadata. EPs call for a new paradigm of research methodology in which digital persistent access to research data are a central issue. In this contribution we highlight 1. The research data as it is archived and curated, 2. the concept "enhanced publication" and its scientific value, 3. the "fragment fitter tool", a language processing tool to facilitate the creation of EPs, 4. IPR issues related to the re-use of the interview data.

1. Introduction

The traditional approach to conducting research in the humanities is to collect dedicated data, add annotations to fragments of the data, analyze these, and publish a paper about the results, crudely said. The data is normally not made available to the scientific public otherwise than in the final publication in which only a fraction of the data appears as an illustration of the researchers' line of reasoning. Re-use of data is however highly desirable for verification and extension of previous research on the data. A good example of re-use of a specific, yet widespread, type of qualitative data, namely interview data is given in the book "Race Talk" (Van den Berg, Wetherell, 2003). The same set of interviews was used by researchers from various disciplines, and the results were published in a book which also contained complete, very detailed transcriptions of the interviews underlying this enterprise. This effort makes the underlying data available for re-use for various disciplines and collective publication of the observations made from the data.

Clearly, interview data can be used in a number of ways, such as comparative research, restudy or follow-up study, re-analysis / secondary analysis, research design and methodological advancement, replication and validation of published work, and for teaching and learning.

For language and speech technology the re-use of data is quite common through central brokerage houses as ELRA and LDC. For interview data in Dutch for example the Spoken Dutch Corpus (Oostdijk, Broeder, 2003) and the IFADV Video dialogue corpus (Van Son, et al., 2008) are available.

However, the general picture for oral history and the social sciences is that a wide range of valuable collections of data are created in specific research projects, but the re-use and secondary analysis of this type of data is obstructed by the fact that most collections are not designed for or available for secondary analysis. There is also a lack of analysis tools for researchers to navigate, analyze and refer to the interview collections. Recent experiences with the re-use of interview data show that there is an enormous potential for this type of data. Especially in the field of interview data related to the Second World War and other military conflicts multidisciplinary research is carried out. The authors are involved in two research projects in which interviews with veterans are used and re-used for qualitative analyses. The first project is called Veteran Tapes VP (VT-VP)[1]. This project resulted in six so-called Enhanced Publications (EPs) about a variety of research questions. These EPs are electronic publications allowing for additional citations in the text being links to fragments of the interviews. By clicking the link the citation becomes available in audio with the associated text (transcription) and meta-data. The second project is called "Living Oral History Workbench" (LOHW), and is funded by the

[1] http://www.surffoundation.nl/en/projecten/Pages/Veteran-Tapes-VP-Enhanced-publication-based-on-multidisciplinary-re-use-of-qualitative-research-files.aspx

Dutch ministry of health, welfare and sports (VWS)[2]. In this project tools are developed to index the interviews with relevant search terms using techniques from Automatic Speech Recognition, and to annotate these in a Wiki-like environment. In this contribution we will address the first project mentioned, Veteran Tapes VP.

We consider this project as exemplary for the paradigm shift that is taking place in the field of Humanities research, more specifically regarding the emerging cyclic character of research in which data curation is important in order to preserve and disseminate data for the long term.

In this paper we will address:

1. The research data used in the project
2. The way the research data is archived and curated
3. The concept "enhanced publication" and its scientific value
4. The "fragment fitter tool", a language processing tool to facilitate the creation of "enhanced publications"
5. IPR issues related to the re-use of interview data.

2. Research Data

The collection used in the project consists of some 25 interviews.This is a subset from a collection of 1.000 interviews compiled by the Dutch Veteran institute[3]. Each interview lasts between 2 and 2.5 hours. In most interviews there are just two speakers, the interview and the interviewee. The interviews are recorded at 44 kHz and downsampled to 16 kHz., PCM encoded wav format. The interviewees were participants in various conflicts and missions in which the Dutch Armed Forces were involved, starting from World War II to most recent peace-missions under the authority of NATO and the UN. The 25 interviews are fully orthographically transliterated including filler marks and indications for speaker such as laughing. The texts are aligned with the audio files in segments of about 3 seconds, containing complete phrases.

The interviews come with a rich body of metadata. The metadata contains personal information about the interviewee (social background, family, education), and the missions s/he served in.

3. Storage of Research Data

In order to facilitate the verification of the research outcomes by others and to enable the further analysis it is important that the interviews are archived in a durable way. The research data will be deposited in the data archive of DANS [4]. DANS – data archiving and networked services – is the Dutch scientific data archive making social sciences and humanities data permanent accessible. The "Dataseal of Approval" (http://www.datasealofapproval.org) contains guidelines for the durable storage of research data such as the interview. An important building block of the archive infrastructure is the "persistent identifier" construct that makes it possible to create persistent links between a scientific article and the research data. This is further explained below as part of the "enriched publication". The interview data will be processed by DANS in order to meet the quality requirements of the CLARIN-NL initiative [5]. The interviews are documented. Specific licenses between DANS and the Veterans Institute arrange the access to sound recordings and transcriptions.

4. Enhanced Publications and their Scientific Value

EPs are envisioned as compound digital objects, which can combine heterogeneous but related web resources. The basis of this compound object is the traditional academic publication. This refers to a textual resource with original work, which is intended for reading by human beings, and which puts forward certain academic claims (Woutersen-Windhouwer, et al., 2009: p.99).

EPs are a new way to publish scientific and other results in an electronic article. The advantage of EPs is that the relation between the article and the underlying data facilitate the peer review process and other quality assessment activities. Due to the link between de publication and the research data the publication can be much richer than a paper edition permits.

Currently, a number of initiatives implement the concept "enhanced publication". An example of a project were EPs were produced by a variety of researchers is the DRIVER II project. A demonstration can be viewed at http://driver2.dans.knaw.nl/.

Also the Veteran Tapes project aims at the realisation of EPs. Based on the available interviews of veterans, a multidisciplinary group of researchers write their research paper in a standard text editor. <TITEL EN EDITORS VAN HET BOEK AL BEKEND???>. Subsequently, an EP version of the paper is written. At passages where the scholars want to include references to the interview material they add links to the data. In traditional publications these references are typically citations from the interviews. In the EPs of The VeteranTapes this will be links to interview fragments. In each link the fragment is displayed in a separate window, showing the transliteration of the fragment, an audio player with the spoken equivalent, and the metadata as selected by the researcher. The fragment can also contain annotations such as references to books or links to websites.

---

2 http://www.minvws.nl/dossiers/erfgoed-van-de-oorlog/projecten/getuigen-verhalen/ (in Dutch)

3 http://interview.veteraneninstituut.nl (in Dutch)

4 http://www.dans.knaw.nl/en

5 http://www.clarin.eu & http://www.clarin.nl

Figure 1 shows an example of a fragment that is linked to an EP.

.

In this way, much more and much richer fragments from the interviews can be included in the publications than is usually presented in a paper version.

Clearly, EPs have much more to offer to the user. However, presently there are clear obstacles for researchers to embrace EPs. Extra effort is needed to add research data to a publication, and although the researchers may see the scientific need to make this effort, at the same time they are confronted with the fact that most A-rated journals do not accommodate for EPs. Since researchers are judged on the impact rate of their publications this poses a serious drawback for making EPs. This means that apart from the researchers also the publishers of highly ranked journals should adopt the EPs as a more valuable way of reporting research. This will require a substantial change in business plans and infrastructures of the publishers.

In The VeteranTapes we have allowed the researchers to write a traditional paper about their findings and have it published in the usual way. This paper is then used as the backbone for the enhanced version of the publication.

5. The fragment fitter tool

We have developed a user-friendly tool to help the researcher linking fragments of the VeteranTapes interviews to the publication.

The idea behind the tool is the following:

- Before using the tool, the researcher preselects the fragments s/he wants to include in the enhanced publication
- The researcher uses the Fragment Fitter to:
  - Find the fragment by indicating the begin time and end time
  - Inspect the corresponding audio segment by listening
  - Anonymise or improve the transcription if needed
  - Add an annotation
  - Select the desired metadata from a list
- After text adjustment the Fitter presents an option to disconnect the audio signal from the fragment
- The researcher links the fragment to the relevant passage in the text
- The fragment becomes visible (and audible) upon clicking the link

We have implemented the tool as an interactive web application. It uses Ajax (JavaScript and XML) at the client side and PHP scripts and a mysql database at the server side. The transcriptions and metadata are stored on the server in mysql database tables and the audio in wave files.

After logging in, the user (researcher or an assistant) selects the interview from a list. From here Figure 2, which is a screen shot from the tool, illustrates the next steps. The application retrieves the interview and shows all the segments of the interview as list of time codes with corresponding transcriptions. After the user has selected a segment, the application cuts the appropriate part of the audio file and makes it available via a player. Also the transcription of the segment is shown in a window. The user may modify the transcription for purposes of correction or anonymisation. In the latter case s/he may also disconnect the audio segment from the fragment. Next, the user can add an annotation to the fragment, e.g. a remark or a reference to publication or website. Finally, the user selects the metadata s/he considers relevant for the readers. Now, the fragment is ready. As a result, the tool generates an XML page with all the information about the fragment (begin & endpoint, corresponding audio segment, transcription, selected metadata). This XML file can be presented in a web page in different ways. In our project it is shown using a php-script we developed. Fig.1 shows a screenshot of the result.

6. IPR issues

As the project is based on the principles of oral history, contrary to the methodology of the social sciences, the interviewees are not anonymous. Moreover, the dataset attached to the interview contains a lot of information on the social and personal background of the veteran. The aim of oral history is to include the narratives of social groups or of experiences that usually remain hidden from history. Researchers who would like to consult the interviews and use excerpts have to sign a form which states that they are obliged to present the material they want to use to the keeper of the collection for clearance. He or she is responsible to respect the law on privacy and to protect first and third parties. Developments in the field of technology now offer possibilities to create similar agreements online such as in the form of licences.

With regard to the Veteran Tapes EPs the following policy is adopted.

Points of Departure:

1. All quotations in an EP should be accessible for every reader
2. The quotations should be duly anonymised
3. Readers/researchers who would like to consult more of the interviews will need the permission to access the interviews as stored at DANS
4. Permission is asked via DANS and granted by the Veteranen Instituut
5. After the permission is granted, access will proceed via authentication through DANS

The veterans have signed a consent form which gives the Veteran Institute the right to use their interview. Their privacy is formally protected by the law on archives and the archives-consultation form. However, they have not given explicit permission to use their voice recordings in

publications. For that reason, we asked the interviewees for permission to use audio portions of the interviews in publications. If they were reluctant to allow this for the full interview, we asked their permission for the specific fragments that were selected for the EPs.

In conclusion, the introduction of a clearance-form which gives permission to differentiated usages of the interview, ranging from the social science-tradition of complete anonymity to the humanities-tradition of a traceable identity, is desrable. This would stimulate the reuse of qualitative data in various forms (such as EP with sound-excerpts).

7. Acknowledgements

This project was made possible by the support of the SURF, the higher education and research partnership organisation for Information and Communications Technology (ICT). For more information about SURF, please visit www.surf.nl.

8. References

Hoogerwerf, M. (2009) Durable enhanced publications. In: *Proceedings of African Digital Scholarship & Curation 2009.*

Oostdijk, N., Broeder, D. (2003). The Spoken Dutch Corpus and its exploitation environment. In: *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03). 14 April, 2003. Budapest, Hungary.*

Van den Berg, H., Wetherell, M., Houtkoop-Steenstra, H. (2003). Analyzing Race Talk. Multidisciplinary approaches to the interview. Cambridge University Press.

Van Son, R., Wesseling, W., Sanders, E., Van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. In: *Proceedimgs LREC 2008, Marrakech, Morocco.*

Woutersen-Windhouwer, S., Brandsma, R., Hogenaar, A., Hoogerwerf, M., Doorenbosch, P., Dürr, E., Ludwig, J., Schmidt, B., Sierman, B. (2009). Enhanced Publications: Linking Publications and Research Data in Digital Repositories. Amsterdam University Press Online available at: http://dare.uva.nl/document/150723

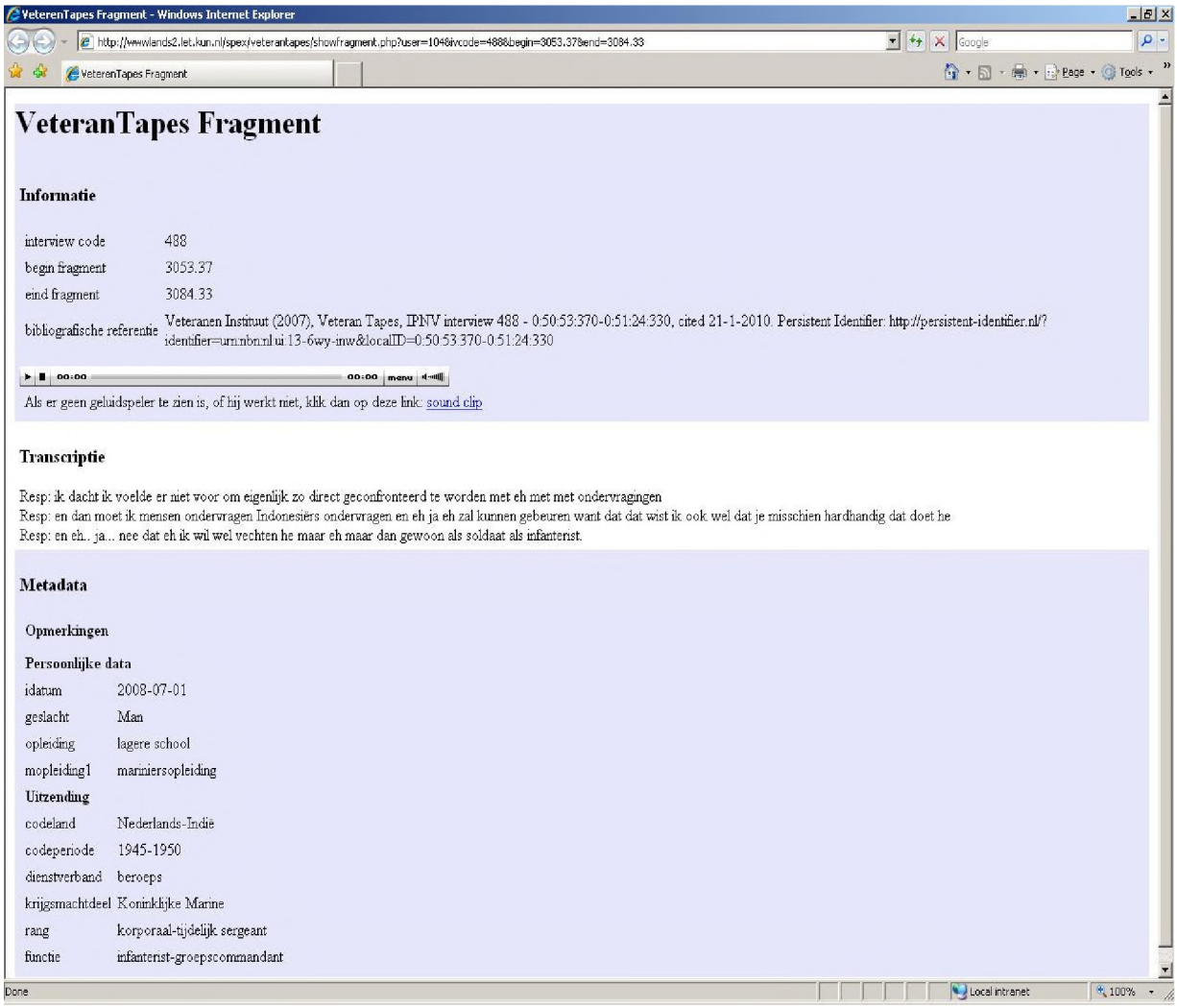

Figure 1: Screenshot of an interview fragment (an enhanced version of a traditional citation)

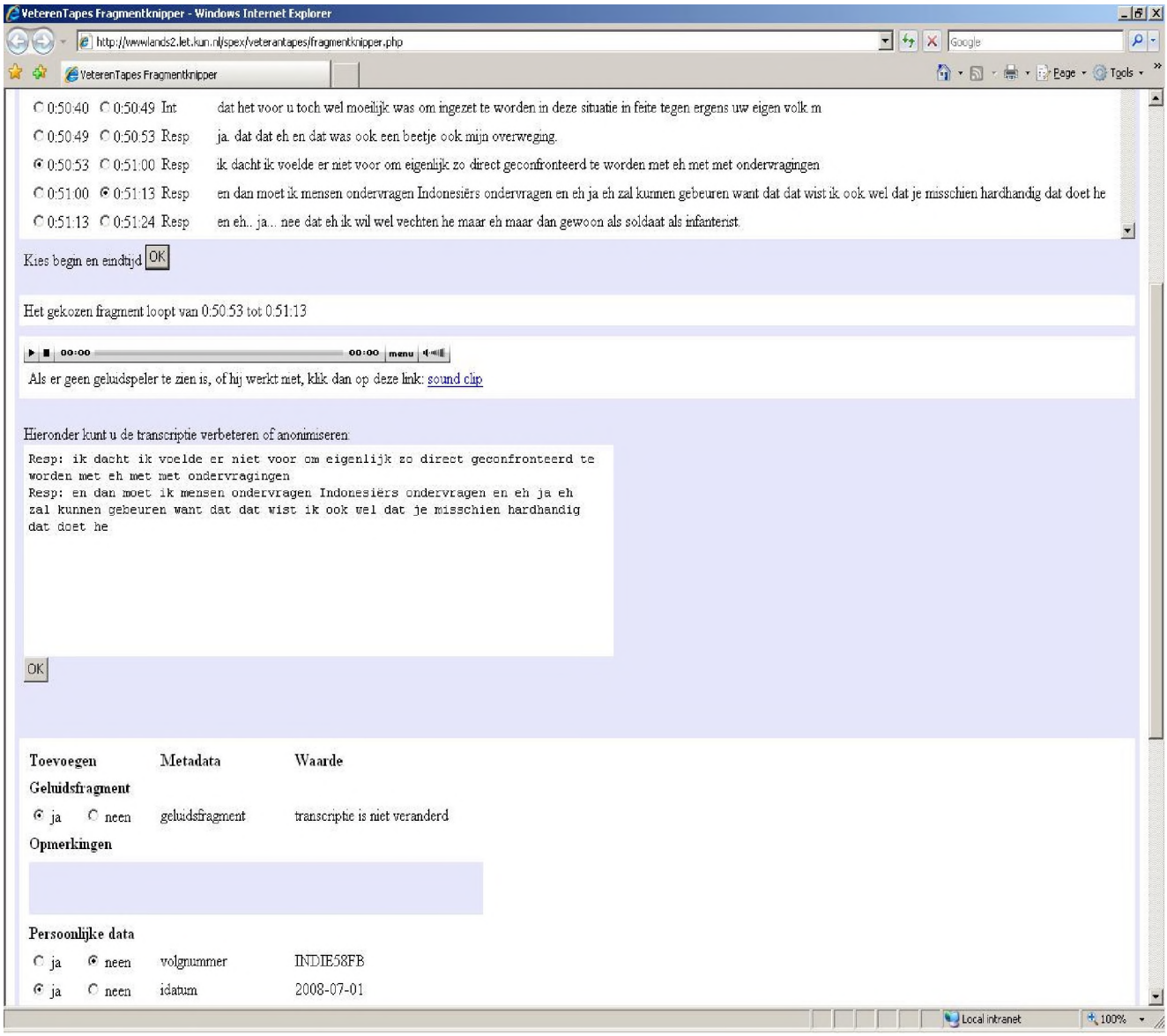

Figure 2: screen shot from the fragment fitter tool