Curation report IPNV

Eric Sanders, Henk van den Heuvel (DCS, CLST, RU Nijmegen), Marjan Grootveld (DANS)
2 augustus 2013

Introduction

The IPNV Corpus is a corpus originally compiled by the Veteraneninstituut (VI). It comprises a collection of more than 1100 (recorded) interviews with veterans who were involved in wars and other military actions that the Dutch military forces took part in. The average duration of an interview is 2.5 hours. Most interviews are with veterans of World War II, the decolonization wars with Indonesia and New Guinea, the UN action in Korea, the UN observe mission in Lebanon, UN missions in Cambodia and former Yugoslavia, and the NATO missions in Iraq and Aghanistan. Some 100 interviews are with veterans who were involved in small-scale observation, monitoring and humanitarian missions.

In the INTER-VIEWs project (CLARIN09-015) 246 of the interviews were curated: the audio recordings (in WAV format) of the interviews were transferred to DANS and the metadata were made available in CMDI/ISOcat format. The data and metadata can be accessed through the EASY system.

For the remaining interviews all recordings are in WAV format as well. They have also been transferred to DANS by the Veteraneninstituut. With these data, some metadata (at least covering Dublin Core categories) is available. The Veteraneninstituut has provided additional metadata (in an MS-Access database) such that the metadata are comparable (and thus compatible) with the metadata for the 246 interviews that were curated in the INTER-VIEWs project.

Here is an overview of the interviews and their public status:
530 Unrestricted
457 Restricted
54 Parked
66 Embargo
=1107 interviews.

Around 950 interviews were curated (including an update of the 246 previously curated interviews). All corresponding cmdi metadatafiles were delivered at DANS. DANS is authorised

to publish various aspects of the metadata in accordance with their agreement (Convenant) with the Veteraneninstituut.

1. Restoring data

In INTER-VIEWs 246 interviews were present with CMDI file. In IPNV a total of 1107 interviews are available in audio. In the metadata MS-Access database file are 1109 entries (interviews) with metadata. For various reasons not all interviews were suitable for curation or publication in DANS EASY:

- 18 interviews are with the interviewers themselves, they were never meant to be included.
- 5 interviews in the metadata file are unknown to DANS.
- 133 interviews do not have a Persistent Identifier (PI), because the information provided by the Veteraneninstituut is insufficient for publication in DANS EASY.

For every interview suited for curation two CMDI files were created (one public, one private, see next section), except for the 18 interviews with interviewers.
The CMDI files are delivered in two sets: one with interviews with PI and one with interviews without PI.

2. CMDI metadata profile & CMDI files

The same metadata profile that was developed in the INTER-VIEWs project was used. A small number of modifications were introduced as compared to the profile of the INTER-VIEWs project:

- Elements without an ISOcat reference were either replaced by one with an ISOcat reference or an ISOcat reference was created for the element.
- The component 'DANS-DC-metadata' was added for metadata harvesting of the CMDI files from the DANS EASY system.
- A few small changes in the minimum or maximum number of occurrences for a metadata category were made.

The CMDI files were constructed with a perl script from the XML-output of an MS Access database that was provided by the company Become-IT. The metadata provided by Become-IT was completely restructured compared to the metadata that was used for INTER-VIEWs, but almost all needed metadata could be found in the new structure. The only element that was not in the new structure was 'onderwerp', but its value was quite redundant anyway. It was also noticed that some values were empty for a couple of interviewees that were not empty in the earlier structure of the metadata, e.g. relationship status and number of children.

Although the profile of IPNV is almost identical to that of INTER-VIEWs, not the same set of components and elements are used in the CMDI files:

- The component InterviewAnnotation is left out the new CMDI's, since there are no usable annotations.
- The component DANS-DC-metadata is added to the new CMDI's, as described above.
- Of some elements the description was slightly changed, e.g. in DistributionMedium the mention of the annotationtool was removed. The keywords and summaries were improved. The formatting of Duration was changed, e.g. from 1.5 hours to 1:30:00
- In all files an anonymous summary of the whole interview is added.

In INTER-VIEWs there were two versions of each CMDI-file, one with private information, meant for researchers who were given permission to see this information and one without private information for the rest of the world. In IPNV there are still two versions of each CMDI-file, both available for everyone. One with private information for interviews of which permission is given by the interviewee to publish private information and one without private information for the other interviews. Only one of the two ends up in the database, which is to be decided by DANS, who has the (dynamically varying) permission information according to their agreement with the Veteraneninstituut.
The decision to change the original dichotomy to the new one was made primarily in consultation with the previous IPNV researcher. This demonstrates that data curation benefits from having access to the original data creator.

The version of the file is reflected in the name:
- <name>_withprivateinfo for files with private information and
- <name>_publicinfo for files without private information.

Because the public/private difference is not anymore related to the person who sees the data, but to the interviewee, the metadata of the interviewer is the same in both files. Items that are in the private, but not in the public CMDI files are:

- Summary of every ten minutes
- Name, social family role and profession of interviewee

The changes in the items that are in the public and private CMDI-files as compared to INTER-VIEWs are:
- In the public version residence place and name of interviewer and interviewee are removed.
- Social family role and profession of interviewee are removed. Birth place of interviewee is added. Summaries of every ten minutes are left out.
- In the private version residence place and social family role of interviewer and interviewee are removed.

3. Restructuring the database

The only restructuring that took place was the contents and the names of the CMDI files as described in the previous section.

4. Converting formats

No files were converted.

5. Documentation

During the project actions and decisions were logged in a Google document.
The important information is in this document.

6. Persistent identifiers

Persistent identifiers were created by DANS and are included in the CMDI files.

7. Transfer data to CLARIN data centre

For the IPNV project the audio files were transferred directly from the Veterans Institute to DANS. The DCS never had the audio files.
On May 31 2031 the CMDI files were made available to DANS by putting the files on a webserver in two zip formats (tar.tgz and 7zip).
The remaining task for DANS was to complete the CMDI files by adding the DANS-DC-metadata component and ResourceProxy element and to make the metadata harvestable for CLARIN-NL.