

Interview Transcription Guidelines and Model Timescapes

23 July 2008 – current

Introduction

The Timescapes Archive Strand has recently finished defining a set of guidelines accompanied by an exemplar which provides a consistent and standard format for transcriptions. This model specifies a level of conformity required for digital preservation while not being unnecessarily prescriptive. The model also attempt to meet diverse requirements: easily explained to transcribers, usable for direct analysis by researchers, readable in hard copy, compatible with qualitative software applications, and suitable (or easily adapted) for web publication and preservation. There are several sections of the exemplar: a cover page that contains project, interviewee and interview details; a header that contains collection and interviewee identifier information and appears on every page; and the body of the transcription.

Cover page, header and body

The cover page provides key identifying information about the collection from which the interview came, such as the name of the project, the depositor, the interviewer and the date of the interview. In addition, any number assigned to the interview is here, as is the name (or pseudonym) of the respondent. Transcriber details and some version history (such as whether or not this transcript has been anonymised) are also provided. This cover page is modeled on guidelines provided by the British Library Sound Archive Oral History section and Timescapes gratefully thanks Rob Perks for providing this information. (The precise fields included are flexible; for more information about the cover page, see the final section on “Metadata”.)

Next, there is header information carried on every page of the transcript. The information includes a project name and an identifier for the interviewee, in this case, a name as permission to use real names has been obtained. This information appears on every page because it is useful for identifying the respondent and connecting an individual interview to the larger collection from which it came. While the practice of printing hard copy is declining, it was still deemed worthwhile to retain this information on every page of a transcript. Use of the header is recommended, but not essential. This header appears only in Microsoft Word versions of documents; once a document is saved as Rich Text Format (.rtf) or plain text (.txt), it no longer appears. A header can, of course, be reinserted into a transcript in an open file format such as an Extensible Markup Language (XML) file by using style sheets.

Finally, some guidelines are provided for the body of the transcript. Regarding formatting content itself, the intent has been to offer minimal guidelines. The goal was to use a format that was easily readable by users, conformed to the input guidelines for most qualitative data software applications and facilitated the process needed for preservation. These guidelines change often as new software versions are released and it is always advisable to review such requirements before finalising a transcript format. One requirement is to use a system of identifying speakers that is flexible enough to

accommodate multiple respondents or focus group participants or interviewers. The template uses “I” for the interviewer, “S” for subject, permitting expansion to more individuals as needed. Timescapes has opted for short tags (one to two characters) in part to permit the showing of the maximum amount of text in a web display and also because conversion programs have been set up to read these characters. Names or initials can also be used. Timescapes also requires the use of speaker tags for its transcripts (other conventions use formatting such as bold for questions, regular text for replies but these should not be used). This is related to computer programs that read transcripts. Perfectly consistent turn-taking can not be assumed (e.g., interjections from third persons do occur). Speaker tags are the easiest way for a computer program to identify the speaker of each utterance.

These guidelines suggest using the traditional colon “:” following a speaker tag. The CAQDAS Networking Project reported that some software packages may not handle the colon. Because the requirements for CAQDAS software change so often, it is suggested that researchers check their applications then decide whether to use colon, hyphen or another punctuation mark.

Clearly, far more detailed specifications can be provided regarding transcription, such as the inclusion or not of non-verbal utterances, background noises, and conventions for transcribing accents, non-standard usage, pauses, grammatical errors, and so on. Such decisions are inextricably linked with the purposes of the research. The intent here has been to leave as much control over such matters to researchers, while still providing a flexible and useful model that meets minimum preservation requirements. There are numerous standards available, Jeffersonian, British Library, and others. It is advisable to follow an existing standard rather than improvising. Within these flexible guidelines, there are some characters to avoid. These are identified in the section on layout and formatting.

Extensible Mark-up Language (XML) and the Text Encoding Initiative (TEI)

Finally, a comment about the use of XML is necessary. Currently, the UK Data Archive uses Rich Text Format as a preservation format for text. However, it is planned that the standard will migrate to XML in the near future. At the moment, XML is not a standard tool for most qualitative researchers. In fact, this is not a big problem as converting from Rich Text Format to XML is not complex and can be handled by short, custom-made computer programs. However, this conversion process is made far easier if a standard layout is followed. Our goal is to allow researchers and transcribers to work in their familiar Microsoft Word (.doc) and Rich Text Format formats, but to do so in a way that ensures conversion to XML later is as easy as possible.

For more detailed mark-up within texts, the UK Data Archive is following the Text Encoding Initiative (TEI) Guidelines. These guidelines are an international standard for representing text in digital form and have been widely used since 1994 in the humanities, linguistics and social science.

We gratefully acknowledge comments and suggestions for this format from many people, with particular thanks to Ann Lewins of the CAQDAS Networking Project, Arja Kuula of the Finnish Social Science Data Archive and Rob Perks of the British Library

Sound Archive. Many members of the Timescapes project also reviewed this model and provided feedback and suggestions based on the requirements of their particular projects.

Layout and formatting

- Formatting, such as bold and italic, will be removed if a file is converted into plain text or XML. Thus the use of such formatting should be avoided if the information needs to be retained in all versions of the transcription. For purposes of clarity in print versions, the UK Data Archive supports the British Library recommendation of Times New Roman font, size 12. However, for purposes of conversion to XML, font and size are at the researchers' discretion.
- Line spacing follows existing conventions of single spacing within a speaker's turn, and double spacing between speakers. As well as being a familiar and readable format, this spacing is also one required for some CAQDAS packages in order to automatically code questions and responses.
- Text is added in brackets [] to indicate start and end points of each side of each audio tape.
- Pagination throughout the document is recommended.
- Line numbering is permitted if done using Microsoft Word's line numbering feature. This feature is easily turned on and off.
- Special characters are generally permitted, e.g., [] for overlapping speech. There are three exceptions: &, < and >. These characters have special meaning for XML and must not be used for general transcription.
- Where annotations are added (adding the word illegible, blank, or other project specific items) ensure this is done consistently. Documentation should include a key or explanation of how these terms are used
- More detailed punctuation is acceptable. What is done in the body of the transcript is, largely, determined by the needs of the research. The template is an example and not intended as a model for quantity of punctuation to be used.

Anonymisation and sensitive content

Transcription often has to consider the issue of whether to indicate confidential or sensitive material that might require anonymisation or other processing at a later date. Timescapes recommends using the following system to indicate sensitive text. At the start of the text to be marked, use the punctuation marks @@@. At the end of the text, use the characters ####. The reason for using these characters is that they are highly unlikely to appear for any other reason in the text. Also, it is important to have different marks at the beginning and end of a segment to prevent the marked and unmarked sections of text from getting out of sequence should the closing tag be forgotten. More detailed guidelines for anonymisation can be found in the document Timescapes Anonymisation Guidelines.

Metadata and creating output for diverse purposes

These guidelines are, in a sense, attempting to straddle a digital divide. On one hand, they are intended to produce a simple, readable transcript with key information retained

with the document. However, they are also intended to create digital texts for publication and preservation. Two cases may be illustrative.

In the first case, researchers are not collecting extensive metadata and have no immediate plans for web publication. They are not using databases or spreadsheets. However, they do follow the recommended format and collect selected metadata on their interview cover pages. In fact, they add to the list of items on the cover page, observing this key principle: keep all fields left justified and followed by a colon. The colon acts as a “delimiter” allowing computer programs to easily locate and copy any text after the punctuation mark. Because the researchers have followed these recommended guidelines, their study metadata can be easily extracted and reprocessed into suitable preservation formats.

In the second instance, researchers have opted to use an Excel spreadsheet (a database would also be acceptable) to hold essential metadata. They do not need detailed data on their cover sheets and have opted to use fewer fields than shown in this model. This is acceptable as the required metadata is to be provided in the Excel spreadsheet. They are able to offer their data in a ready-for-preservation structure: interviews are in Rich Text Format with essential identifiers provided and all additional metadata is in the spreadsheet. Computer programs that it is hoped will eventually be available for public use take the transcripts and metadata as inputs and use style sheets to produce various layouts, such as one version that will be used to download data and another that will be for web display.

As with all qualitative methodology, there are almost no invariant rules and context is critical. If you have questions or comments about these guidelines, please contact Libby Bishop at e.l.bishop@leeds.ac.uk.

*Sample transcript cover page – data are a mix of real and fictional***Project Information:**

Programme name: Pioneers
Project name: Family Life and Work Experience Before 1918
Depositor name: Paul Thompson

Interviewee Information:

ID number: 2000int002
Title: Mr
First name: John
Surname: Stinchombe
Date of birth: 1886
Gender: Male
Marital status: Married
Occupation: Farm Labourer

Interview Information

Interviewer first name: Paul
Interviewer surname: Thompson
Transcriber first name: Jane
Transcriber surname: Smith
Anonymisation: Permission granted to use real names-P. Thompson 2007
Place of interview: Colchester
Date of interview: 23 February 1976
Number of tapes/files: 2
Type of recorder: Marantz Model 007 cassette audio recorder
Consent status: Consent form signed for participation and archiving
Keywords: fatherhood, siblings, turning points, historical events

[Start of Tape 1, Side A]

- I: If I could ask you when you were born?
- S: 1896.
- I: Whereabouts was this?
- S: Over at Horton, Woodley Arton.
- I: Is this a village or is it -?
- S: Oh no, I was out on the wilds as you might say. I had the farthest for to go to school. Two mile to walk.
- I: Had you any other houses round about?
- S: Well the farmer as I worked for thirty years. I was born in a cottage about 300 yards away as the crow flies. And I went to live in the old house where, as he bought afterwards. When the {ILLEGIBLE} estate was sold up. The farmers, they worked hard before that see. One of his brothers lived in that house and he was born there. And his uncle lived there before that. And then there was another brother down at his mill farm. Well he down at mill farm he was more boss than the one that lived in this old thatched house.
- I: What was your father's job?
- S: Farm labourer. Like meself.
- I: It was the farmer's cottage that you had then?
- S: That's only a cottage on the farm. That's the last cottage in the lane. In them days.
- I: How large was it?
- S: What the cottage? Well this was two up and two down. Then where I went to live, after the war, before I got married, cos I was married in 1915, and I went to this other old farm house in 1918. And then we stopped in there until 1948 when I shifted over to Old Sodbury on account of the wife cos if she wanted anything extra we had to come to Sodbury for to get it. And that was no unthankful office especially on a bike, a push bike. When you get out on Sodbury Common you know which way the wind is blowing. And I used to come across there of an evening if we wanted summat, extra. Sometimes I have had to stand on the pedals and I had to get off and walk when the wind was so rough all against me.
- I: Yes, bad isn't it.

S: Oh it is.

I: Had you a garden at your first cottage where you were brought up?

S: Oh that was two gardens as you might say. One was one side of the road a three cornered piece and the other was adjoining the cottage. That wasn't what, {blank} oh I suppose all together it wouldn't be half an acre. Then when I went to live in this old thatched farm house well that was half an acre of garden to say nothing about the lawn and all that. And twelve month before we left over there I had the first prize for the garden. And he was twelve perch long. And wasn't the full length of the garden as was dug up see. And that was 96 rows of 'taters, two feet apart. And sometimes when the onions and all that during the war, well I didn't plant, the biggest part were onions and the rest of em were shallots. And sell 'em. For to get the extra shilling.

I: Did your father grow vegetables as well?

S: Oh yes.

I: What sort of things did he grow?

S: Some of them all sorts.

I: I see. Had you any fruit?

S: In this other, in this other old thatched house, it used to be a farmhouse in years gone by. Well I put it down to 300 year old but I was a long way out as it was more than 1600.

I: Yes, very old.

S: That was six rooms on him, three up and three down. Terrible great beams they were anyone wouldn't realize it.

S: And when we went into it, I had to put some thatch on first because the estate lot see let it go. And I put some on there because the thatcher that used to do all the houses thatching he was away at timber cutting. When he came home well the farmer had it thatched all over see then. He done it once afterwards. While I was there. When I left there, the fellow that took my job see, both of them was a little bit touched, I am sorry to say, and the old doctor said afterwards when the master, he came and see the missus, Mrs Stinchcombe, {BLANK} imagine that, and {ILLEGIBLE} they come from the beans. It was a darned great big whitened passage and {BLANK} just inside the front door and then there was the baking oven adjoining that see.

[End of Tape 1, Side A]