Journal of the Text Encoding Initiative

Issue 15 | 2024
Selected Papers from the 2021 TEI Conference

InTEIrviews: An ODD for Qualitative Interviews in the Humanities

Marie Puren and Florian Cafiero

Electronic version
URL: https://journals.openedition.org/jtei/5007
DOI: 10.4000/jtei.5007
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Marie Puren and Florian Cafiero, "InTEIrviews: An ODD for Qualitative Interviews in the Humanities", *Journal of the Text Encoding Initiative* [Online], Issue 15 | 2024, Online since 24 April 2024, connection on 29 November 2024. URL: http://journals.openedition.org/jtei/5007 ; DOI: https://doi.org/10.4000/jtei.5007

InTEIrviews: An ODD for Qualitative Interviews in the Humanities

Marie Puren and Florian Cafiero

ABSTRACT

Qualitative interviews are a crucial research method for many disciplines in the humanities and social sciences. Yet apart from rare initiatives, interview transcriptions are seldom shared with other researchers. Even when they are, annotations are mostly absent, and rarely standardized or designed to be read by anyone but the initial research team. In this paper, we advocate for a more open management of these resources, and present a proposition for an XML-TEI–conformant standard that allows for their accurate transcription and annotation. The ODD we present is aimed at facilitating systematic analyses of corpora of interview transcriptions, as well as ensuring wider dissemination and reusability of these resources. Relying entirely on existing TEI elements to create this ODD, we combine primarily the elements and attributes declared by the modules "Transcription of Speech" and "Language Corpora."

Journal of the Text Encoding Initiative, Issue 15, 24/04/2024
Selected Papers from the 2021 TEI Conference

This ODD is designed to help describe the actual development of the interview, transcribe the interpretations made by the researcher(s), and tackle ethical or anonymity issues.

INDEX

Keywords: qualitative interviews, standardization, transcription of speech, spoken discourse

ACKNOWLEDGEMENTS

We are thankful to participants at the International Conference for Computational Social Science (IC2S2 -2021 / Massachusetts Institute of Technology) and at the Next-Gen TEI Conference 2021 for their useful comments, especially to Lou Burnard, James Cummings, Kathryn Tomasek, and Joe Wincentowski. Previous discussions with Emilien Schultz were also extremely helpful to formalize this project. Errors remain our own.

1. Introduction

1 Qualitative interviews—that is, in-depth, open-ended conversations with participants to explore their thoughts, experiences, beliefs, or perceptions—constitute an important research tool for disciplines such as history (Smith 2003), sociology (Lamont and Swidler 2014), ethnography (Skinner et al. 2013), political science (Mosley 2013), and applied linguistics (Talmy 2010). By engaging in discussion with a series of subjects, the researchers learn about both facts and the meaning of these facts for members of a certain category of interest (Plas and Kvale 1996).

2 As intuitive as it may seem, interviewing is a complex and costly process. Choosing and contacting participants, designing an interview guide, running the interviews and properly recording, transcribing, and then analyzing them—all this takes dedication, time, and financial resources. Making the most of this material is thus crucial. Encoding the interviews and their annotations properly can help to share them with reviewers or colleagues, increasing transparency and trust in the method (Corti, Fielding, and Bishop 2016). It also can help one to reuse the interviews years after for another purpose, or simply to be less emotionally invested (Mauthner, Parry, and Backett-Milburn 1998). Finally, it can help researchers who did not collect the data to work on this material for *secondary analysis*, a more and more common practice generating an increasing number of

publications (Heaton 2008; Hughes and Tarrant 2019). Yet despite rare initiatives such as beQuali (Cadorel et al. 2018), the Finnish Social Science Data Archive, or the UK Data Service (Bishop and Kuula-Luumi 2017), interview transcriptions in social sciences are very rarely shared with other researchers. Even when the data are made available, their annotation is most of the time conceived only for personal use, without following any sort of standard or explicit rules. A few oral history projects, such as the pioneering work of *Voice of the Holocaust*,[1] or the recent tools developed by the Dartmouth Digital History Initiative, have led the way in using XML-TEI (DDHI, n.d.). The latter, for instance, proposed to encode speakers, places, persons, organizations, dates, and events in TEI.

3 In this paper we advocate for more open management of these resources and present a proposition for an XML-TEI–conformant standard that allows for their accurate transcription and annotation.

2. Why use TEI for qualitative interviews?

4 TEI is a world unto itself, and a certain investment is required to master its principles. One can legitimately ask why one should use TEI to encode qualitative interview transcripts when there are software programs that support the transcription and annotation of interviews, and when a word processor seems to be sufficient to meet the majority of interviewers' needs. Moreover, TEI may seem inordinately complex for researchers who may fear that its adoption will generate more problems than benefits for their research project (Romary 2009). But a number of arguments can be made to show the value of using TEI, especially in the context of open and replicable science.

2.1 A community with a bottom-up approach to digital data management

5 The Text Encoding Initiative aims to provide a framework for creating and managing all data in digital form produced or used by researchers in the humanities and social sciences. As the term *Initiative* implies, it brings together individuals and organizations interested in the problems surrounding the encoding of these data in electronic form. The strength of the TEI is that it is above all a community that brings together all kinds of practitioners of text as a source of study or as data to be enriched, analyzed, and visualized with a computer. It is therefore not only an enterprise of technicians, but from the start a project that brings together computer scientists and humanities and social-science scholars (Romary 2009). The TEI advocates not a *top-down* but a *bottom-up* approach: it is indeed a community of practice, coupled with knowledge of the

"field," that has given rise to these guidelines. The TEI community therefore offers a particularly favorable environment for proposing solutions that facilitate the processing, analysis, and sharing of interview data. The aim here is to respond to the demands and needs of a specific community: qualitative interviews are used by numerous disciplines, each of them using different types of interview in different ways and for different purposes. Even within a single discipline such as sociology, "the term applies to an enormously wide range of research practice" (Lamont and Swidler 2014), each with its own requirements.

2.2 Producing FAIR interview data with TEI

6 Established in 2014, the FAIR principles (Wilkinson et al. 2016) have been increasingly successful, and have been adopted more widely by researchers, governments, policy-makers, and funding bodies seeking guidelines for managing and disseminating sustainable and reusable scholarly data (Tóth-Czifra 2019). The fifteen FAIR (Findable Accessible Interoperable Reusable) principles consist of a set of minimum recommendations designed to facilitate the discovery, access, and reuse of data by data providers and consumers, whether machines or humans (Wilkinson et al. 2016). One of the reasons for the success of these principles is that there are no technical prerequisites for implementing them.

2.2.1 A widely adopted standard

7 To strengthen the interoperability and reusability of interview data, it is necessary to consider using standards. With the TEI Guidelines, two people working in different environments can use the same elements and attributes to encode the same types of textual phenomena. This promotes the creation of reusable data. TEI has all the characteristics that identify it as a de facto standard for the representation of texts in digital form (TEI Consortium 2021): namely that it expresses consensus, has public and easily accessible documentation, and is actively maintained (Romary 2011). TEI remains a particularly open format—"born to be open," as Laurent Romary describes it (Romary 2020)—as it offers users the possibility of adapting it to their needs. As mentioned above, TEI is based on a "bottom-up" approach, enabling the scientific community not only constantly to evolve its guidelines, but also to make free use of the various elements at its disposal.

8 While TEI offers the possibility of markup perfectly adapted to the needs of a specific project or community, this high level of granularity can also diminish the interoperability of the data produced, and therefore their potential reuse. For this reason, it is necessary to propose specific, documented TEI use cases that best cover the specific needs of the scientific communities identified, and thus offer a shared, evolving framework for creating digital data within this subcommunity.

2.2.2 Embedded metadata

9 The interoperability and reusability of data is also ensured by the addition of descriptive metadata. Metadata are indeed essential to the future reuse of digital data (Gilliland 2016). When one has access to an interview conducted and transcribed by another, it is in fact necessary to have easy access to descriptive information on the interview's conditions of production, lest one risk misinterpreting the data or having to work, without knowing it, with incomplete data (Brunel et al. 2021).[2]

10 Metadata, when they exist, are often stored and managed separately. This can be problematic, as one can *lose* this metadata file, or no longer know exactly what resource it describes. TEI addresses this problem with the mandatory `<teiHeader>` element, which allows the metadata attached to a document to be represented, while the document's textual content is represented in the `<text>` element. As a container for the metadata, the TEI header stores all this information in one place, regardless of how it may be used. It provides information about the TEI file, but also about the source that is being transcribed, or about the production conditions of the digital file that is being encoded. For while the `<teiHeader>` has a bibliographic function, "it [also] provides a place for everything" (Burnard 2014), allowing it to meet the needs of a wide variety of research communities. Furthermore, the mandatory presence of the `<teiHeader>` in the TEI document has another positive effect: it serves as a reminder of the need to produce metadata that describes precisely how the interview was conducted. Hopefully, the use of TEI will encourage transcribers to systematically document the interviews they produce.

3. How to use TEI to annotate a qualitative interview?

3.1 Customize the TEI to fit the transcription of qualitative interviews

11 At the root of TEI we find a certain universalistic impulse, since TEI is intended to cover all fields of textual production as long as the text *exists* in electronic form. Since its beginning, the creation and evolution of TEI has been strongly related to the development of digital humanities (Burnard 2012), and thus to the representation of ancient texts, but it is gradually evolving into a conceptual framework for encoding all types of texts (Romary 2009), and it can be used for any type of data in electronic form (Burnard 2014), such as transcriptions of speech.

12 To facilitate the use of TEI by interview data producers, we propose to create an XML schema suited to the needs of qualitative interview practitioners. The ability to validate an XML document against the rules of the language and its structure is one of the many benefits of XML, as it helps to ensure that information exchanged between agencies conforms to the agreed-upon formats. As Lou Burnard points out, however, it is particularly rare that anyone needs all the elements of TEI (Burnard 2014). This is especially so for interview transcripts, where the encoding requires mainly elements that can encode the turns of speech between the interviewee(s) and the interviewer or annotate the events specific to this type of exchange.

13 The use of a schema specifically adapted to qualitative interviews has two main benefits. On the one hand, it has the advantage of providing a "turnkey" solution for those who are not specialists in XML-TEI or scholarly editing. By limiting the number of elements and attributes that can be used, the schema makes it possible to limit the available choices, and thus facilitate the use of TEI. The creation of such a schema does not prejudge its evolution: if we hope to propose here a list of elements well chosen enough to satisfy most of the needs of qualitative interview transcribers, we are also aware of the need to accommodate those who wish to modify this schema, by either adding new elements and attributes or deleting others. The power of TEI lies in this flexibility, which allows it to be customized according to one's needs. On the other hand, a schema for transcribing qualitative interviews would facilitate the emergence of a standard for creating such data. Following Martin Holmes and Laurent Romary's work on a TEI schema for scholarly literature, such a schema would strike a balance between (1) prescription, which encourages encoders to

adopt recommendations endorsed by the community; (2) arbitration, which favors an approach adapted to specific needs, with a view to simplicity, interoperability, and uniformity; and (3) codification, which merely formalizes transcribers' current practices (Holmes and Romary 2010).

14 To promote the use of such a schema, we have created an ODD, or "One Document Does it All," which documents the choices made to create this schema in human-readable language (Burnard 2013; Burnard and Rahtz 2004). Offering an XML schema is indeed not enough: one must be able to reach the community it is intended for and acculturate this community to its use. It is therefore necessary to offer a *user manual* that explains the interest of such a schema and how to use it in one's research work, while illustrating this with concrete examples. It seemed to us that the provision of an ODD was fully in line with these objectives. This ODD and the schema it documents are available online on Github.[3]

3.2 The structure of the ODD "InTEIrviews"

3.2.1 Distinguishing between macro- and microstructure

15 We have seen that conducting a qualitative interview is a common practice in the humanities and is carried out by researchers from various disciplines. Hence there are variations in the conventions surrounding transcriptions. As Thomas Schmidt points out, it is particularly important to ask whether these differences in practice may reflect "pure idiosyncrasy," or whether they may also mark fundamental differences in research aims or theoretical perspectives (Schmidt 2011). It can be argued that the structure of an interview transcript varies little: the description of the conditions of the interview (participants, place, date, etc.) accompanies the transcribed interview, which can be divided into successive turns of speech, sometimes interrupted by incidents (noise, distracted participants, etc.). On the other hand, transcription reflects the research objectives, and the conventions used must therefore be able to meet these objectives.

16 Substantial latitude must be left to the transcriber, who needs to be able to annotate a wide range of phenomena. Following Thomas Schmidt, we distinguish two levels of standardization: the macrostructure and the microstructure of the interview. The macrostructure can be represented in a generic TEI format based on TEI P5; the representation of the microstructure is based on more

fine-grained TEI markup. The advantage of using TEI is that it allows for preserving interview-specific variations while expressing them in a standardized way, and thus facilitates the processing of these transcripts.

17 Thanks to a customization of TEI for transcribing qualitative interviews, it is possible to cover the main characteristics of this textual genre with the TEI Guidelines, and to represent both the macrostructure of a qualitative interview and its microstructure, defined as "the low-level component of the full-text content" (Holmes and Romary 2010). The macrostructure of an interview in its transcribed form is as follows:

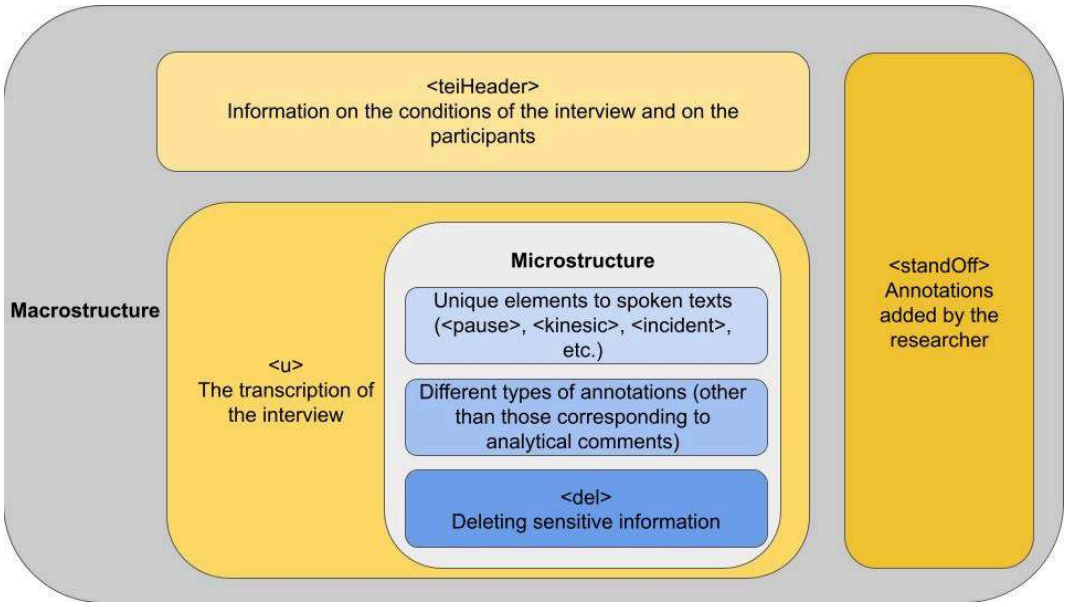
- Information on the conditions of the interview and on the participants. This information is contained in the `<teiHeader>`.
- The transcription of the interview itself, with questions and answers, possibly divided into different parts. This involves encoding the questions and answers using the element `<u>`, accompanied by a `@who` attribute to indicate who is speaking.
- Annotations by the researcher, who will generally include an additional semantic layer with comments. The researcher who conducted the interview and usually transcribes it will comment on the transcript with notes on various aspects. We propose to use the element `<standOff>` to contain these annotations.

18 The microstructure of the interview concerns the "low-level" annotations. Thomas Schmidt sees microstructure as everything that has to do with the form and semantics of the textual elements themselves: it is the "names for, representation of, and relations between linguistic transcription entities like words, pauses, and semi-lexical entities" (Schmidt 2011). The microstructure corresponds in particular to the transcriptional conventions of spoken discourse, that is, the annotation of semantic and syntactic analyses, as well as interferences (e.g., pauses). But as Thomas Schmidt points out, these conventions are numerous, even countless. Based on this definition, we propose the following typology, which attempts to distinguish the main categories of elements that make up the microstructure of qualitative interviews:

- Elements unique to spoken texts that describe external interferences and behaviors of interviewees, such as `<pause>`, `<kinesic>`, or `<incident>`.

- Different types of annotation (other than those corresponding to analytical comments added by the researcher): annotations of named entities (with elements from the "Names, Dates, People, and Places" module) or of various linguistic phenomena (TEI elements related to linguistic annotations, such as `<w>` or `<phr>`). This general category covers different types of annotation, corresponding to disciplinary practices or the needs expressed by the researcher.
- Ethics and privacy. To disseminate qualitative interviews, it is essential to withhold certain information, notably the identity of the participants and any information that might allow them to be identified. It involves deleting passages, and sometimes replacing them. In this context, the use of the `<gap>` element might be a good option. This is used to indicate that part of the transcribed text has been omitted, and the reason why. (See section Ethics and respect of privacy for more on this aspect).

Figure 1. Macrostructure and microstructure of a qualitative interview in XML-TEI.



3.2.2 InTEIrviews

3.2.2.1 Creating the ODD: Combining the "Transcription of Speech" and "Language Corpora" modules

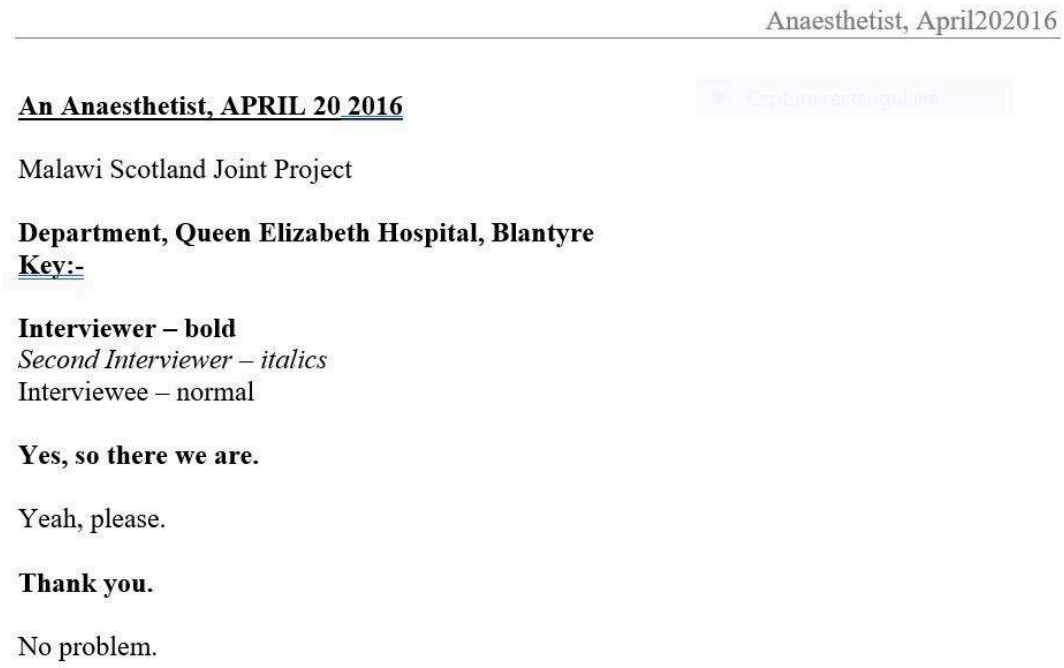
19 To create this ODD, we first turned to the "spoken" or "Transcribed Speech" module, dedicated to the transcription of spoken discourse, which seemed particularly appropriate for the transcription of qualitative interviews. This module was originally designed for linguistic analysis, however, and lacks some of the elements necessary to accurately represent the textual phenomena of and information related to qualitative interview transcription. But a close look at the guidelines shows that elements not belonging to the "spoken" module can be particularly useful. For example, in the "corpus" or "Metadata for Language Corpora" module, the `<particDesc>` element for "description of participation" allows for the encoding of a vital piece of information, namely who the participants in the interview are; indeed, `<particDesc>` "describes speakers, voices, or other identifiable participants in any type of text or other persons named or otherwise referred to in text, edit, or metadata."

20 For this reason, the ODD "InTEIrviews" combines elements from both the "Transcription of Speech" and "Language Corpora" modules. The list of elements thus chosen allows us to encode information about the conditions in which the interview took place, all the phenomena specific to this exercise, and the additional annotations likely to be used in the context of the analysis of qualitative interviews. We have thus tried to personalize the TEI by respecting a double injunction: to restrict the number of selected elements as much as possible while proposing elements adapted to the variety of transcription conventions.

21 We ourselves experimented with these principles on qualitative interviews author 2 had to analyze, regarding the attitudes of mothers towards vaccines in underprivileged neighborhoods in France. Unfortunately, we cannot share the results of our TEI encoding here because of the terms of confidentiality agreed on at the time between the funder of the study and the interviewees. Therefore, in the examples below, we use the transcript of an interview available online in a research data repository (here the *Edinburgh DataShare*). This is an actual qualitative interview conducted as part of a study on outsourcing of external development assistance in maternal and child health (MCH) in Malawi and Nepal.[4] The transcript is in the form of a Word document, entitled "Malawi-KII-MSAPApril202016_anonymized.doc."[5] The document provides some metadata on the conditions of the interview (location and participants), but metadata describing the wider context

of the project are available only online.[6] So the metadata are "scattered," and not directly related to the interview. Using the schema we have created, it is possible to record all this information in the `<teiHeader>`.

Figure 2. Extract from the transcript in Word format.



Example 1. Information contained in `<fileDesc>` and `<encodingDesc>`.

```
<fileDesc>
 <titleStmt>
  <title>An Anaesthetist, APRIL 20 2016</title>
 </titleStmt>
 <publicationStmt>
  <authority>School of Social and Political Science, Edinburgh</authority>
  <availability>
   <licence>Creative Commons License: Attribution 4.0 International </licence>
  </availability>
 </publicationStmt>
 <sourceDesc>
  <recordingStmt>
   <recording>
    <date> 20 April 2016 </date>
```

```
     </recording>
    </recordingStmt>
   </sourceDesc>
  </fileDesc>
  <encodingDesc>
   <projectDesc>
    <p>
     <bibl>
      <title>Interview data from “new norms and forms of development,” 2014-2016
[dataset]</title>
      <author>Adhikari, Radha</author>
      <author>Mandambwe, Khumbo</author>
      <author>Smith, Pam</author>
      <author>Sharma, Jeevan</author>
      <author>Harper, Ian</author>
      <author>Malata, Address</author>
      <author>Chand, Obindra</author>
      <author>Thapa, Deepak</author>
      <date>2017</date>
      <publisher>School of Social and Political Science</publisher>
      <pubPlace>Edinburgh</pubPlace>
      <availability>
       <licence>Creative Commons License: Attribution 4.0 International </licence>
      </availability>
      <idno type="DOI">10.7488/ds/2048</idno>
     </bibl>
     <desc>This data contains findings of the study on outsourcing of external
development assistance in maternal and child health (MCH) in Malawi and Nepal. It
outlines the institutional modalities and norms guiding the financing and delivery
of MCH projects and programs. First, our study of external development assistance
reveals a messy assemblage of actors, institutional arrangements, and activities
informed by the norms “value for money” and “measurable results.” Second, we found
that for development assistance to function effectively it is not just about the
flow of financial resources to a project or a program but also about networks
and key personal and institutional relationships. Third, we found that there is
increasing political pressure to show that the disbursement of resources is linked
to the achievement of measurable results. </desc>
    </p>
```

```
  </projectDesc>
 </encodingDesc>
```

Example 2. Information on the setting and the participants in `<profileDesc>`.

```
 <profileDesc>
  <langUsage>
   <language ident="en"/>
  </langUsage>
  <settingDesc>
   <setting>
    <locale>Department</locale>
    <name>Queen Elizabeth Hospital</name>
    <placeName>Blantyre</placeName>
   </setting>
   <place/>
  </settingDesc>
  <particDesc>
   <listPerson>
    <person xml:id="interviewee">
     <occupation>Anaesthetist</occupation>
    </person>
    <person xml:id="interv1">
     <occupation>Interviewer</occupation>
    </person>
    <person xml:id="interv2">
     <occupation>Interviewer</occupation>
    </person>
   </listPerson>
  </particDesc>
 </profileDesc>
```

3.2.2.2 Some examples of annotations

22 In this section we present some annotation mechanisms, enabled by InTEIrviews, which we believe are particularly pertinent to qualitative interviews.

3.2.2.2.1 Ethics and respect of privacy

23 A key concern in sharing qualitative interviews should be the respect of legal constraints and ethical principles. The aim is to ensure the protection of interviewees while concealing as little relevant information as possible. To share qualitative interviews (either online, e.g., in a data warehouse, or directly with third-party researchers), it may indeed be necessary to withhold personal information, for example, or to remove offensive language. This can happen in several ways:

- You permanently delete the passage concerned.
- You permanently delete the passage concerned, but you wish to replace it (for example with a pseudonym in the case of a person's name)
- You wish to have two versions of the transcribed interview: one containing the deleted passages, and another version that can be disseminated without showing them. In this case, by indicating the passages to be deleted with an XML element, you could take advantage of the possibilities offered by XSLT (eXtensible Stylesheet Language Transformations) to easily and automatically delete the passages concerned.

24 In any case, the transcriber wants to keep track of this deletion. For example, in the interview transcript we are working on, the transcriber has systematically indicated when a person's name has been removed by putting "[*name has been taken out from here*]" between square brackets.

25 To indicate that the transcriber's interventions have deleted parts of the transcribed interview, we propose to use the element `<gap>`, which signals the omission while leaving transcribers free to explain why they omitted these passages. The `<gap>` element is used to indicate that a passage has been omitted by transcribers or encoders "where copy text is not transcribed due to editorial policy or because it is impossible to do so" (TEI Consortium 2021). The strategy for omitting such passages (including in what cases they have been deleted, and what information should be provided about the deletion) should therefore be described in the TEI Header.

26 The attribute `@resp`, borne by `<gap>`, will be used to indicate who is responsible for this intervention; it is indeed useful to know who took the decision to delete this or that passage: the researcher in charge of the project, the transcriber, the person in charge of data dissemination, etc. To give details of the extent of the deleted passage, the attributes `@unit` and `@quantity` may be

used to specify the length of the deletion. The reason for the deletion could then be expressed with the attribute @reason. An element <desc>, enclosed within the <gap> element, can also give more information on the causes of the deletion (legal reasons, ethics code, personal moral judgment, etc.).

Example 3. Use of <gap> to indicate that a passage has been omitted for anonymization purposes.

```
<u who="#interv1">Yes, so there we are. </u>
<u who="#interviewee">Yeah, please.</u>
<u who="#interv1">Thank you,
 <gap reason="anonymized">
  <desc>Name omission</desc>
 </gap>
and it's very good to see you and...</u>
<u who="#interviewee">Oh yeah! <vocal>
  <desc>Laughs</desc>
 </vocal></u>
```

3.2.2.2.2 Reflecting on one's interview practices

27 The <u> element is used to encode the different turns of speech by the interviewers and the interviewees, with a @who attribute to identify the speaker. But qualitative interviews are not ordinary conversations: they are prepared by a researcher, implementing a strategy to get as much information as possible on a topic of interest. It is thus crucial to encode the researcher's comments on his own speech (Beaud 1996): Was the question prepared? Spontaneous? What was its purpose (changing the subject/knowing more/confirming a previous statement, etc.)? This is why we propose to use an attribute @ana, borne by <u>, to describe this kind of information. Concerning the value of this attribute @ana, it could be interesting to limit it to a controlled vocabulary, in order to make practices more easily comparable among researchers, surveys, and projects. In the example below, we show how the attribute @ana can be used. The attribute values were chosen by the article's authors and are for illustrative purposes.

Example 4. Use of the @ana attribute to express the purpose of the interviewer's questions.

```
<u who="#interv1" ana="question_prepared">But that was one question I~was going
to ask you, do you actually have the statistics that show the improvement in
women's and babies' survival?</u>
<u who="#interviewee">That's what I've been doing last week. </u>
```

```
  <u who="#interv1" ana="expression_agreement">Yeah. Oh!</u>
  <u who="#interviewee">I just collected the data and I've already sent it to
Scotland and we want to have a look at it and see whether it's making sense or
not. What I~have seen is that surely... This is a fact, the maternal deaths are
actually going down. Sometimes down to zero in a month.</u>
  <u who="#interv1" ana="relaunch">In any particular...?</u>
  <u who="#interviewee">Yeah, in all these three.</u>
```

3.2.2.2.3 Annotating the interview: Sharing one's interpretation

28 In addition to simple content annotations (persons or places cited, dates evoked, etc.), our model offers the possibility of sharing one's interpretations of relevant passages of the transcription. We propose to use the element `<seg>`, bearing an attribute `@xml:id` to delimit the parts of the speech that are in need of further analysis. This analysis can then be provided via a `<span>` element, with a `@target` attribute, to identify which `<seg>` element is concerned, and a `@type` attribute to express the nature of the analysis. The element `<interpr>` may be used in conjunction with the `<span>` element, but `<interpr>` is more suitable for identifying various parts of dialogue under unique conceptual categories. Associating identified parts of speeches and specific conceptual categories is easy with TEI pointer mechanisms: for example, an `@ana` attribute borne by a `<seg>` element enables associating this element with an `<interp>` element bearing an `@xml:id` attribute. In the example below, we show how to use `<span>` and `<interp>` to add analytical comments. These (fictitious) comments were added by the article's authors to allow them more easily to illustrate the mechanism at work. We use the element `<standOff>` to contain all of these annotations that add information to the transcript—and therefore do not belong to the transcript itself.

Example 5. Use of `<span>` or `<interp>` to annotate the interview.

```
  <span type="comment" target="#an_qeh1">significant investment by doctors
in relation to the resources available</span><standOff><span type="comment"
target="#an_qeh1">significant investment by doctors in relation to the
resources available</span><spanGrp type="topic"><span target="#an_qeh2">Money</
span><span target="#an_qeh3">Training</span></spanGrp><interpGrp><interp
xml:id="funding">Funding problems and mecanisms</interp></interpGrp></standOff>
  [...]
  <u who="#interviewee"><seg xml:id="an_qeh1" ana="#funding ">It's extremely
cheap but bringing in people from other places, yes, that's expensive because
they'll need accommodation and stuff like that</seg> and this is why, just last
```

```
week I~went around to the same hospitals again, I~was trying to collect more data
before so that we can actually say, this is the actual impact of the courses.
[...] </u><u who="#interv1"> And he resuscitated the patient? </u><u who="#
interviewee"> The patient was okay and I~said, wow! <seg xml:id="an_qeh2">So
you can see that people are really doing the impossibles</seg>, you know,
so I~ think it's really a lifesaving course and I~think it's very... <seg
xml:id="an_qeh3">It's very very important that they continue this training</seg>
so that new people coming into the hospital should be actually trained [...]. </u>
```

4. Conclusion

29 This ODD is of course only a first step, but it has already been presented with success to researchers in social sciences and the humanities. The main obstacle to the adoption of such an ODD—and thus of the related XML schema—that we propose is mainly technical. As Scagliola et al. 2020 point out, researchers producing interview data are "willing to integrate a digital tool into their existing research practice and methodological mindset, if it can easily be used or adapted to their needs." The investment required to master new digital tools is crucial to facilitating their adoption. The time needed to master and use XML-TEI can therefore be a barrier to producing transcripts in this format. To promote the use of our ODD, two avenues should be considered:

- Training researchers in TEI early in their careers, and thus communicating more widely with research communities that are not traditionally trained in TEI;
- Facilitating the use of XML-TEI with annotation tools that are easy to use and that do not involve directly modifying the source code.

30 These will be explored in further work. Furthermore, as the practice of transcribing oral discourse is widely shared by the humanities and social sciences community, and therefore by many TEI practitioners, we believe it would be particularly appropriate to consider the creation of a special interest group on transcription. This would contribute to the development of our ODD proposal, its dissemination, and, we hope, its wide adoption by the humanities and social sciences community.

BIBLIOGRAPHY

Beaud, Stéphane. 1996. “L’usage de l’entretien en sciences sociales: Plaidoyer pour l’entretien ethnographique.” *Revue des sciences sociales du politique* 9: 226–57.

Bishop, Libby, and Arja Kuula-Luumi. 2017. “Revisiting Qualitative Data Reuse: A Decade On.” Sage Open 7 (1) https://doi.org/10.1177/2158244016685136.

Brunel, Valentin, Paul Colin, Alina Danciu, Quentin Gallis, and Emilie Groshens. 2021. *Des métadonnées et paradonnées pour mieux exploiter des données d’enquêtes pour la recherche*. Mate la Science ouverte. Virtuel, France.

Burnard, Lou. 2012. “Du literary and linguistic computing aux digital humanities: Retour sur 40 ans de relations entre sciences humaines et informatique.” In *Read/Write Book 2: Une introduction aux humanités numériques*. OpenEdition Press. https://books.openedition.org/oep/226?lang=en.

———. 2013. “The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure.” *Journal of the Text Encoding Initiative*, 5. https://journals.openedition.org/jtei/811 https://doi.org/10.4000/jtei.811

———. 2014. *What is the Text Encoding Initiative? How to Add Intelligent Markup to Digital Resources.* OpenEdition Press. https://books.openedition.org/oep/679?lang=en.

Burnard, Lou, and Sebastian Rahtz. 2004. “RelaxNG with Son of ODD.” In *Proceedings of Extreme Markup Languages*. Montréal: Extreme Markup Languages.

Cadorel, Sarah, Guillaume Garcia, Emilie Fromont, Emilie Groshens, Emeline Juillard, and Jérémie Vandenbunder. 2018. “beQuali: Une plateforme d’archives numériques en sciences sociales.” *Proceedings of the 1st International Conference on Digital Tools & Uses Congress*, 1–5.

Corti, Louise, Nigel Fielding, and Libby Bishop. 2016. “Editorial for Special Edition, Digital Representations: Re-using and Publishing Digital Qualitative Data.” Sage Open 6 (4): 2158244016678911.

DDHI (Dartmouth Digital History Initiative). n.d. “DDHI Encoding Guidelines.” https://ddhi.dartmouth.edu/ddhi-encoding-guidelines.

Gilliland, Anne J. 2016. “Setting the Stage.” In *Introduction to Metadata*, edited by Murtha Baca. Los Angeles: Getty Publications. https://www.getty.edu/publications/intrometadata/setting-the-stage/.

Heaton, Janet. 2008. “Secondary Analysis of Qualitative Data: An Overview.” *Historical Social Research/ Historische Sozialforschung*, 33–45.

Holmes, Martin, and Laurent Romary. 2010. "Encoding Models for Scholarly Literature." Ioannis Iglezakis, Tatiana-Eleni Synodinou, Sarantos Kapidakis. *Publishing and digital libraries: Legal and organizational issues*, pp. 88–110. https://arxiv.org/pdf/0906.0675.

Hughes, Kahryn, and Anna Tarrant. 2019. *Qualitative Secondary Analysis*. Sage.

Lamont, Michèle, and Ann Swidler. 2014. "Methodological Pluralism and the Possibilities and Limits of Interviewing." *Qualitative Sociology* 37 (2): 153–71.

Mauthner, Natasha S., Odette Parry, and Kathryn Backett-Milburn. 1998. "The Data Are Out There, Or Are They? Implications for Archiving and Revisiting Qualitative Data." *Sociology* 32 (4): 733–45.

Mosley, Layna. 2013. "'Just Talk to People'? Interviews in Contemporary Political Science." In *Interview Research in Political Science*, edited by Layna Mosley, 1–28. Ithaca, NY: Cornell University Press.

Plas, Jeanne M., and Steinar Kvale. 1996. "Interviews: An Introduction to Qualitative Research Interviewing." Sage.

Romary, Laurent. 2009. "Questions and Answers for TEI Newcomers." *Jahrbuch für Computerphilologie*, 10. https://hal.science/hal-00348372v2/document.

———. 2011. "Stabilizing Knowledge through Standards: A Perspective for the Humanities." In *Going Digital: Evolutionary and Revolutionary Aspects of Digitization*, edited by Karl Grandin. Center for History of Science at the Royal Swedish Academy of Sciences. https://arxiv.org/abs/1011.0519.

———. 2020. "TEI guidelines: born to be open." ACDH-CH : Austrian Centre for Digital Humanities and Cultural Heritage Lectures, Jun 2020, Vienne, Austria. , Lecture (6.1). https://inria.hal.science/hal-02864525/

Scagliola, Stefania, Louise Corti, Silvia Calamai, Norah Karrouche, Jeannine Beeken, Arjan Van Hessen, Cristoph Draxler, and Khiet Truong. 2020. "Henk van den Heuvel." *Cross Disciplinary Overtures with Interview Data: Integrating Digital Practices and Tools in the Scholarly Workflow*. https://doi.org/10.3384/ecp2020172015.

Schmidt, Thomas. 2011. "A TEI-based Approach to Standardising Spoken Language Transcription." *Journal of the Text Encoding Initiative*, vol. 1. https://journals.openedition.org/jtei/142 https://doi.org/10.4000/jtei.142.

Skinner, Jonathan, ed. 2013. *The interview: An Ethnographic Approach*. ASA Monographs, vol. 49. A&C Black.

Smith, Richard Cándida, et al. 2003. "Analytic Strategies for Oral History Interviews." In *Inside Interviewing: New Lenses, New Concerns*, p. 347. https://doi.org/10.4135/9781412984492.

Talmy, Steven. 2010. "Qualitative Interviews in Applied Linguistics: From Research Instrument to Social Practice." *Annual Review of Applied Linguistics* 30: 128–48.

TEI Consortium. 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* [Version 4.3.0, modified 31 August 2021]. https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/index.html.

Tóth-Czifra, Erzsébet. 2019. "The Risk of Losing the Thick Description: Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem." In *Digital Technology and the Practices of Humanities Research*, 235–66. OpenBook Publishers. https://doi.org/10.11647/OBP.0192.10.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

NOTES

1 https://voices.library.iit.edu/.

2 For example, an interview is not always transcribed in full.

3 https://github.com/mpuren/InTEIrviews.

4 The data set can be downloaded at https://datashare.ed.ac.uk/handle/10283/2713.

5 The complete Word document can be found at https://github.com/mpuren/InTEIrviews/blob/main/Malawi-KII-MSAPApril202016_anonymized.doc.

6 The complete metadata can be found at https://datashare.ed.ac.uk/handle/10283/2713?show=full.

AUTHORS

MARIE PUREN

Marie Puren is an associate professor in History and Digital Humanities in the LRE laboratory at EPITA Paris. She is the head of the MNSHS team (Digital Methods for Humanities and Social Sciences). She is also an associate researcher at the Centre Jean Mabillon of the École nationale des chartes—PSL (Paris, France). She works on the creation of processing chains for digitized historical sources, and on the annotation and analysis of textual data extracted from these documents. She is interested in the management of FAIR data for the humanities.

FLORIAN CAFIERO

Florian Cafiero is a fellow in Artificial Intelligence applied to the humanities and social sciences at the Paris Sciences et Lettres University (Paris, France) and a researcher at the Centre Jean Mabillon of the École nationale des chartes—PSL. His lectures focus on digital humanities (École nationale des chartes—PSL, École normale supérieure de Paris—PSL) and computational social science (Université Paris-Dauphine—PSL).