Transcription: representing texts in space and time

TEI@Oxford

2010-07

Transcription

Which features of a primary source might one want to include in a transcription?

- variant letter forms
- page layout
- orthography
- capitalisation
- word division
- punctuation
- abbreviations
- additions and deletions
- errors and omissions

But wait! We also transcribe spoken texts and much of this applies to them as well!

Elements defined for transcription

Defined in 'core' module: <abbr>, <add>, <choice>, <corr>, <del>, <expan>, <gap>, <sic>

Defined in 'transcr' module: <addSpan>, <am>, <damage>, <damageSpan>, <delSpan>, <ex>, <facsimile>, <fw>, <handNotes>, <handShift>, <restore><space>, <subst>, <supplied>, <surface>, <zone>

Using <choice> in transcriptions

- <choice> (groups alternative editorial encodings)
- Abbreviation:
  - <abbr> (abbreviated form)
  - <expan> (expanded form)
- Errors:
  - <sic> (apparent error)
  - <corr> (corrected error)
- Regularisation/normalisation:
  - <orig> (original form)
  - <reg> (regularised form)

Abbreviation and expansion

An abbreviation may be transcribed in two ways:

- One may choose to give the unexpanded abbreviation, transcribing it simply as a particular sequence of letters or marks on the page: thus, a ‘p with a bar through the descender’ or an ‘a with a macron’
- One may also interpret or ‘expand’ the abbreviation, supplying the letter or letters it is seen as standing for: thus, ‘per’, ‘an'

The TEI allows one to provide both the abbreviated and expanded forms.

Encoding abbreviations

The TEI proposes two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion can be encoded using <abbr> and <expan>
- the mark or sign used to indicate the suppression of one or more letters, and the letters supplied in the process of expansion can be encoded using <am> and <ex>

Here too one may also use both levels simultaneously.

<am> and <ex>

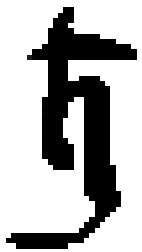
Using these elements, from the 'transcr' module, the transcriber may indicate the status of the individual letters or signs within both the abbreviation and the expansion.

- <ex> (editorial expansion) contains a sequence of letters added by an editor or transcriber when expanding an abbreviation.
- <am> (abbreviation marker) contains a sequence of letters or signs present in an abbreviation which are omitted or replaced in the expanded form of the abbreviation.

Previously, people have re-purposed existing elements such as <hi> and <supplied> to mark individual letters/signs in abbreviations and expansions. The new P5 elements <am> and <ex> are the TEI's attempt to support this desire.

A simple example

The Icelandic word ‘hann’ (‘he’) is frequently written in medieval manuscripts as the letter ‘h’ with a horizontal stroke or bar (Unicode character 0305, functionally similar to the modern tilde). It looks like this:



Encoding abbreviations 1

Depending on editorial policy, we might represent this in any one of the following ways:

```
<abbr>h&#x305;</abbr> or
<expan>hann</expan>
```

```
 h<am>&#x305;</am>or h
<ex>ann</ex>
```

```
<abbr>h<am>&#x305;</am>
</abbr> or
<expan>h<ex>ann</ex>
</expan>
```

Using <choice>

Any of these pairs can be wrapped in <choice> tags:

```
 h<choice>
  <am>&#x305;</am>
  <ex>ann</ex>
</choice>
```

```
<choice>
  <abbr>h<am>&#x305;</am>
  </abbr>
  <expan>h<ex>ann</ex>
  </expan>
</choice>
```

Corrections and emendations

The <sic> element can be used to indicate that the reading of the manuscript is erroneous or nonsensical, while <corr> (correction) can be used to provide what in the editor's opinion is the correct reading:

```
<sic>blowe</sic>
```

```
<corr>blow</corr>
```

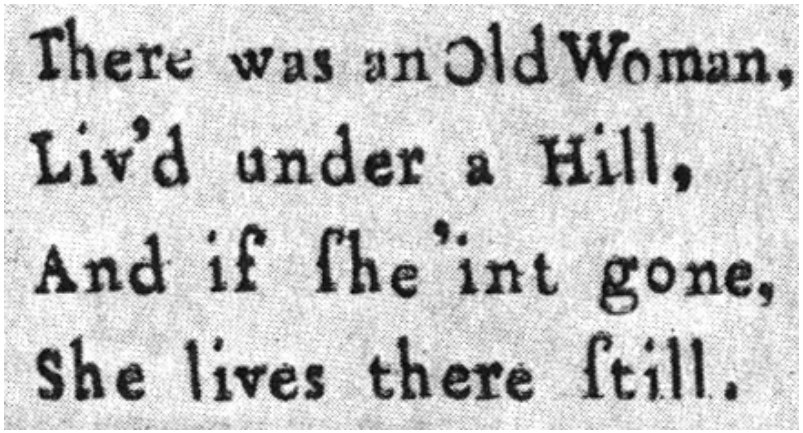
The two may be combined within a <choice> element:

```
<choice>
  <sic>blowe</sic>
  <corr>blow</corr>
</choice>
```

Normalisation/regularisation

Source texts rarely use modern normalised orthography. For retrieval and other processing reasons, such information may be useful in a transcription. The <reg> (regularized) element is available used to mark a normalised form, while the <orig> (original) element indicates a non-standard spelling. These elements can optionally be grouped as alternatives using the <choice> element.

Normalisation/regularisation (example)




```
<lg>
  <l>There was an Old Woman,</l>
  <l>
    <choice>
      <orig>Liv'd</orig>
      <reg>Lived</reg>
    </choice> under a hill,</l>
  <l>And if she <choice>
      <orig>'int</orig>
      <reg>isn't</reg>
    </choice> gone,</l>
  <l>She lives there still.</l>
</lg>
```

Additions, deletions and substitutions

Alterations made to the text, whether by the scribe or in some later hand, can be encoded using <add> (addition) or <del> (deletion).

- <add> (addition) contains letters, words, or phrases inserted in the text by an author, scribe, annotator, or corrector.
- <del> (deletion) contains a letter, word, or passage deleted, marked as deleted, or otherwise indicated as superfluous or spurious in the copy text by an author, scribe, annotator, or corrector.

Where the addition and deletion are regarded as a single *substitution*, they can be grouped together using the <subst> (substitution) element .

Substitutions

<subst> (substitution) groups one or more deletions with one or more additions when the combination is to be regarded as a single intervention in the text. Examples:

- one word/letter written over another
- one word/letter deleted, replaced by another written above it by the same hand at one time
- one word/letter deleted, replaced by a different hand some other time
- a long chain of substitutions on the one stretch of text, with uncertainty as to the order of substitution and as to which of many possible readings should be preferred

<add> and <del> Examples

```
<l>In Flanders fields the <subst>
   <del>poppies</del>
   <add>flowers</add>
  </subst>
  <subst>
   <del>blow</del>
   <add>grow</add>
  </subst>
</l>
```

<add> and <del> Examples (2)

```
<l>Take up our <subst>
   <del>quarrel</del>
   <add>
    <subst>
     <del>fight</del>
     <add>
      <choice>
       <sic>quarell</sic>
       <corr>quarrel</corr>
      </choice>
     </add>
    </subst>
   </add>
  </subst> with the foe:</l>
```

<addSpan> and <delSpan>

These two elements delimit a span of text by pointing mechanisms rather than by enclosing it. This is useful if an addition or deletion overlaps another span of text.

@spanTo indicates the end of a span initiated by the element bearing this attribute.

```
<addSpan spanTo="#id4"/>
<!-- added text -->
<anchor xml:id="id4"/>
```

Cancellation of deletions and other markings

<restore> indicates restoration of text to an earlier state by cancellation of an editorial or authorial marking or instruction.

If in the line 'For I hate this my body' from D.H. Lawrence's poem *Eloi, Eloi, Lama Sabachthani?*, the 'my' was first deleted then restored by writing 'stet' in the margin, this might be encoded thus:

```
For I hate this
<restore hand="#dhl" type="marginalStetNote">
  <del>my</del>
</restore> body
```

Text omitted from or supplied in the transcription

Where a word has been supplied by the editor, <supplied> can be used. It is customary to distinguish between text now illegible or lost through damage but assumed originally to have been in the manuscript (which in some editorial traditions is printed in square brackets), and text assumed to have been inadvertently omitted by the scribe (printed in angle brackets). This distinction is indicated in the mark-up through the use of the *@reason* attribute:

```
…Dragging the worst
among<supplied reason="omitted">s</supplied>t us…
```

Metadata for supplied text

Attributes *@resp* and *@cert* can be used here as elsewhere. A *@source* attribute is also available to indicate that another witness supports the reconstruction:

```
<p>ath þeir
<supplied reason="omitted" source="AM02-152" cert="high">mundu</supplied>
sundr ganga</p>
```

When missing text cannot be confidently reconstructed, the <gap> element should be used. Its *@reason* attribute explains the reason for the omission and its *@extent* and *@unit* attributes indicate its presumed size.

```
<gap reason="damage" extent="7" unit="chars"/>
```

Other uses of <gap>

The <gap> element can also be used where material present and legible has been omitted in a transcription, whether for editorial reasons or as part of sampling practice.

```
<div rend="slide">
  <head>Lectio x.</head>
  <p> Hic itaque paterfamilias ad excolendam
<gap extent="20" unit="words" reason="not transcribed" resp="#DC"/>
congregare non desistit.
  </p>
</div>
```

Damage and illegibility

Use <unclear> if the text has been rendered partly illegible by deletion or damage so that the text can be read but without perfect confidence.

Use the *@reason* attribute to state the cause (damage, deletion etc.) of the uncertainty in transcription and the *@cert* attribute to indicate the confidence in the transcription.

```
 shore of the <unclear reason="damage" cert="medium">the Hudson,
at</unclear> that broad
```

The <damage> element should be used to indicate areas of damage affecting the text, but normally where at least some of the text can be read with confidence. The attributes *@agent* and *@extent* indicate the cause and extent of the damage respectively.

<handNote> and <handShift>

The <handNote> element is used to provide information about each hand distinguished within the encoded document.

- When the 'transcr' module is used, the element <handNotes> is available, within the <profileDesc> element of the Header, to hold one or more <handNote> elements. (brief)
- When the 'msdescription' module is included, the <handDesc> element also becomes available as part of a structured manuscript description. (more robust)

It is possible to use the two elements together if, for example, the <handDesc> element contains a single summary describing all the hands discursively, while the <handNotes> element gives specific details of each.

<handShift>

<handShift> marks the beginning of a sequence of text written in a new hand, or the beginning of a scribal stint.

```
<l>When wolde the cat dwelle in his ynne</l>
<handShift medium="greenish-ink"/>
<l>And if the cattes skynne be slyk <handShift medium="black-ink"/> and
gaye</l>
```

```
<handNotes>
  <handNote xml:id="h1" script="copperplate">Carefully written with
    regular descenders</handNote>
  <handNote xml:id="h2" medium="pencil">Unschooled scrawl</handNote>
</handNotes>
```

<handShift> example

```
<handShift new="#h1" resp="#das"/>... and that
good Order Decency and regular worship may be once
more introduced and Established in this Parish
according to the Rules and Ceremonies of the
Church of England and as under a good
Consciencious and sober Curate there would and
ought to be <handShift new="#h2" resp="#das"/> and
for that purpose the parishioners pray
```

@hand, @resp, @cert

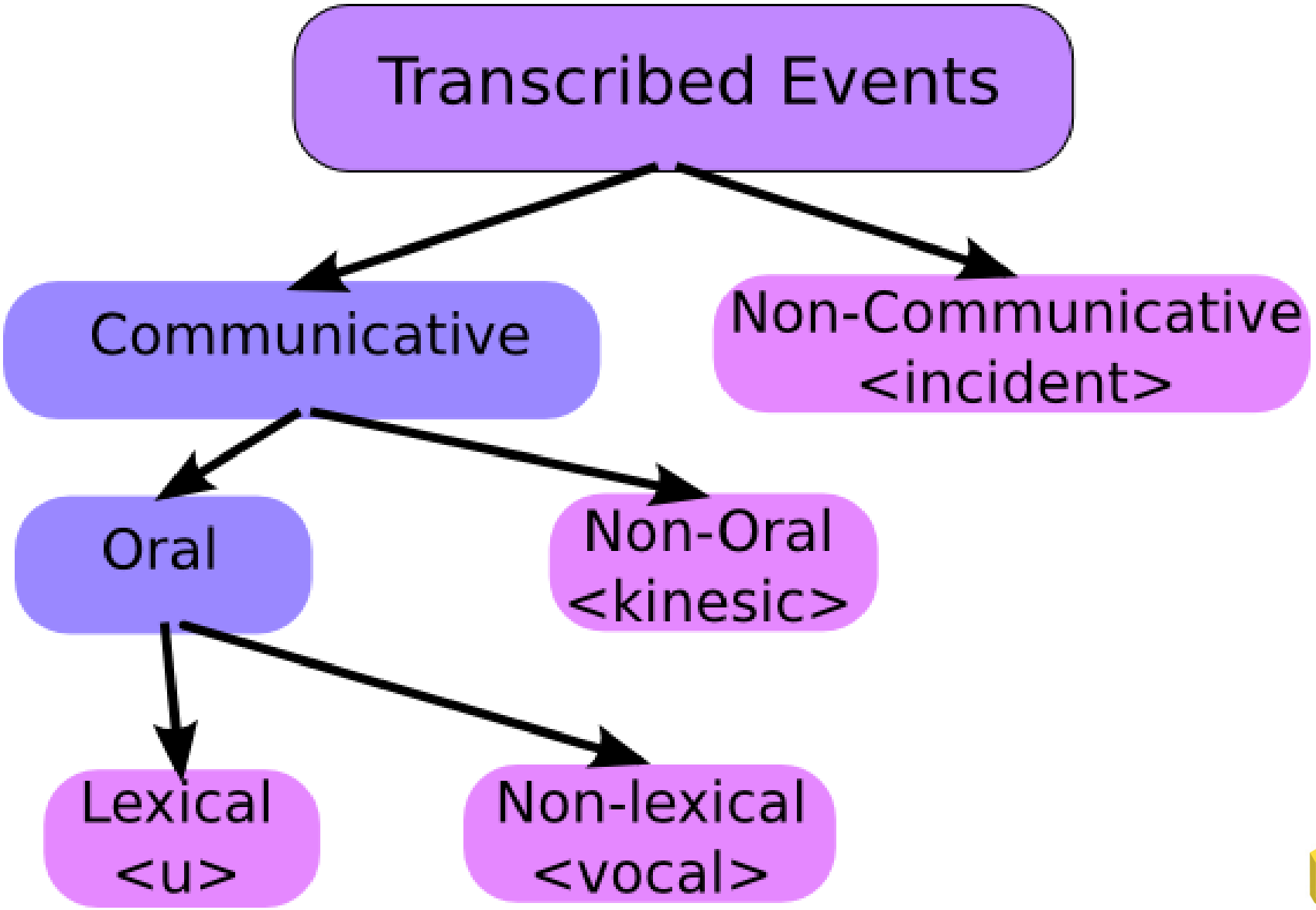
```
<add place="supra" hand="#WJ" cert="medium"> But</add>
<choice>
  <sic>One</sic>
  <corr resp="#FB" cert="high">one</corr>
</choice> must have lived ...
<!-- elsewhere -->
<respStmt xml:id="FB">
  <resp>editorial changes</resp>
  <name>Fredson Bowers</name>
</respStmt>
<respStmt xml:id="WJ">
  <resp>authorial changes</resp>
  <name>William James</name>
</respStmt>
```

Spoken Texts

A spoken text may contain any of the following components:

- utterances
- pauses
- vocalized but non-lexical phenomena such as coughs
- kinesic (non-verbal, non-lexical) phenomena such as gestures
- entirely non-linguistic incidents occurring during and possibly influencing the course of speech
- writing, regarded as a special class of incident in that it can be transcribed, for example captions or overheads displayed during a lecture
- shifts or changes in vocal quality

What sort of events?



<teiCorpus> reminder

Grouping documents into a corpus allows you to factor out the metadata they have in common:

```
<teiCorpus>
  <teiHeader>
<!-- shared metadata -->
  </teiHeader>
  <TEI>
    <teiHeader>
<!-- specific metadata -->
    </teiHeader>
    <text>
<!-- ... -->
    </text>
  </TEI>
  <TEI>
    <teiHeader>
<!-- specific metadata -->
    </teiHeader>
    <text>
<!-- ... -->
    </text>
  </TEI>
</teiCorpus>
```

The notion of "utterance"

- problematic, but pragmatic
- a sequence of speech from a single speaker
- may be grouped into higher-level s
- or fragmented into smaller segments <seg> or <s>
- the @*who* attribute points to speaker information

Transcriptions of Speech

Elements defined: <broadcast>, <equipment>, <incident>, <kinesic>, <pause>, <recording>, <recordingStmt>, <scriptStmt>, <shift>, <u>, <vocal>, <writing>,

Classes defined: att.duration, model.divPart.spoken, model.global.spoken, model.recordingPart

Simple examples

Mixture of utterance and 'paralinguistic' information:

```
<u who="#Jan">This is just delicious</u>
<incident>
  <desc>telephone rings</desc>
</incident>
<u who="#Kim">I'll get it</u>
<u who="#Tom">I used to <vocal>
    <desc>coughs</desc>
  </vocal> smoke a lot</u>
<u who="#Bob">
  <vocal>
    <desc>sniffs</desc>
  </vocal>He thinks he's tough
</u>
<vocal who="#Ann">
  <desc>snorts</desc>
</vocal>
<u who="#Tom">Yeah
<kinesic>
    <desc>gives uplifted middle finger sign</desc>
  </kinesic>
</u>
```

Back channelling

```
<u who="#a">So what could I have done <vocal who="#b">
   <desc>tut-tutting</desc>
  </vocal> about it anyway?</u>
```

Example using other TEI elements

```
<u who="#mar">you never <pause/> take this cat for
show and tell
<pause/> meow meow</u>
<u who="#ros">yeah well I dont want to</u>
<incident>
  <desc>toy cat has bell in tail which continues
     to make a tinkling sound</desc>
</incident>
<u who="#ros">because it is so old</u>
<u who="#mar">how <choice>
    <orig>bout</orig>
    <reg>about</reg>
  </choice>
  <emph>your</emph> cat <pause/>yours is <emph>new</emph>
  <kinesic>
    <desc>shows Father the cat</desc>
  </kinesic>
</u>
<u trans="pause" who="#fat">thats <pause/> darling</u>
<u who="#mar">no <emph>mine</emph> isnt old
mine is just um a little dirty</u>
```

Shifts in voice quality

- Classic multiple hierarchy problem
  - can use <shift> or <milestone> to mark boundaries...
  - ... or can use typed <seg> elements
- useful also for code shifting

```
<u who="#LB">
  <shift feature="loud" new="f"/>Elizabeth
</u>
<u who="#EB">Yes</u>
<u who="#LB">
  <shift feature="loud"/>Come and try this <pause/>
  <shift feature="loud" new="ff"/>come on!

</u>
```

Sample prosodic feature list

(based on Boase, Survey of English Usage, 1990)

tempo	(fast, slow, getting faster, slower, etc.)
loud	loud, soft, getting louder, slower
pitch range	high, low, wide, narrow, ascending...
tension	slurred, tense, staccato, legato...
rhythm	regular, irregular, spiky rising or falling...
voice quality	whisper, husky, falsetto, giggle, sobbing, yawning, sighing...

Researchers need to define their own terms

<shift/> vs <incident>

Compare:

```
<u who="#a">Listen to this <shift new="reading"/>The government is
confident, he said, that the current economic problems will be completely
overcome by June<shift/> what nonsense!</u>
```

and

```
<u who="#a">Listen to this
<incident>
   <desc>reads aloud from newspaper</desc>
 </incident> what nonsense!</u>
```

<vocal> vs <u>

Compare:

```
<vocal who="#ann">
  <desc>snorts</desc>
</vocal>
```

and

```
<u who="#ann">
  <vocal>
    <desc>snorts</desc>
  </vocal>
</u>
```

<writing> example

```
<u who="#a">look at this</u>
<writing who="#a" type="newspaper" gradual="false">
Government claims economic problems <soCalled>over by June</soCalled>
</writing>
<u who="#a">what nonsense!</u>
```

Timing issues

- pausing: use <pause> element
- duration: use @*dur* attribute
- synchronization: use @*synch* attribute
- overlap: use @*trans* attribute

```
<u>Okay <pause dur="PT2M"/>U-m<pause dur="PT75S"/>the scene opens up
<pause dur="PT50S"/> with <pause dur="PT20S"/> um <pause dur="PT145S"/>
you see a tree okay?</u>
```

Overlap

```
Mutt: Have you heard the --
Jeff: the election result?
Mutt: It's a disaster!
```

```
<u who="#mutt">have you heard the</u>
<u trans="latching" who="#jeff">the election result</u>
<u who="#mutt">its a disaster</u>
<u who="#jeff" trans="overlap">its a miracle</u>
```

Synchronization

```
<u who="#mutt">have you heard <anchor synch="#t1"/>the</u>
<u who="#jeff" synch="#t1">the election result</u>
<u who="#mutt" synch="#t2">its a disaster</u>
<u who="#jeff" synch="#t2">its a miracle</u>
<!-- Elsewhere in Document -->
<timeline origin="#t1">
  <when xml:id="t1"/>
  <when xml:id="t2"/>
</timeline>
```

Using Elements Seen Elsewhere...

```
<u>
  <del type="truncation">s</del>see
<del type="repetition">you you</del> you know
<del type="falseStart">it's</del> he's crazy

</u>
```

```
<gap reason="passing truck" extent="5" unit="s"/>
```

```
<u who="#P1">I proposed that <foreign xml:lang="de"> wir können
<pause dur="PT1S"/> vielleicht </foreign> go to warsaw and
<emph>vienna</emph>
</u>
```

Metadata: Participant Description

```
<particDesc>
 <listPerson>
  <person xml:id="P-1234" sex="2" age="mid">
   <p>Female informant, well-educated, born in Shropshire UK, 12 Jan 1950,
of unknown occupation. Speaks French fluently. Socio-Economic status
B2.</p>
  </person>
  <person xml:id="P-4332" sex="1">
   <persName>
    <surname>Hancock</surname>
    <forename>Antony</forename>
    <forename>Aloysius</forename>
    <forename>St John</forename>
   </persName>
   <residence notAfter="1959">
    <address>
     <street>Railway Cuttings</street>
     <settlement>East Cheam</settlement>
    </address>
   </residence>
   <occupation>comedian</occupation>
  </person>
 </listPerson>
</particDesc>
```

Metadata: <scriptStmt> Example

```
<sourceDesc>
  <scriptStmt xml:id="CNN12">
    <bibl>
      <author>CNN Network News</author>
      <title>News headlines</title>
      <date when="1991-06-12">12 Jun 91</date>
    </bibl>
  </scriptStmt>
</sourceDesc>
```

Metadata: <recordingStmt> Example

```
<recordingStmt>
  <recording type="audio" dur="P30M">
    <respStmt>
      <resp>Location recording by</resp>
      <orgName>Sound Services Ltd.</orgName>
    </respStmt>
    <equipment>
      <p>Multiple close microphones mixed down to stereo Digital Audio Tape,
standard play, 44.1 KHz sampling frequency</p>
    </equipment>
    <date>12 Jan 1987</date>
  </recording>
</recordingStmt>
```

Detailed <recording>

```
<recording type="audio" dur="P10M">
  <equipment>
    <p>Recorded from FM Radio to digital tape</p>
  </equipment>
  <broadcast>
    <bibl>
      <title>Interview on foreign policy</title>
      <author>BBC Radio 5</author>
      <respStmt>
        <resp>interviewer</resp>
        <name>Robin Day</name>
      </respStmt>
      <respStmt>
        <resp>interviewee</resp>
        <name>Margaret Thatcher</name>
      </respStmt>
    </bibl>
  </broadcast>
</recording>
```

Conclusions

- The TEI provides detailed recommendations for transcription of primary sources, even when these sources are spoken texts, audio, or video that we've only touched on briefly here.
- Detailed acts of intervention (scribal or editorial) can be indicated through the hierarchy of XMl elements.
- 'Transcriptional' and other elements can be used in conjunction with those for transcribing spoken texts.
- The TEI is also adopting a community proposal for 'Genetic Editing' which records in much more detail the marks on a document and the stages of textual revision leading to the creation of a text. We have a workshop on this today!