Digital editing of primary sources: an overview of the TEI proposals

Lou Burnard Consulting
Digital editing ... in practice

A scholarly digital edition may have any or all of the following components:

- a set of digital images, each representing a page (or other surface) of some source
- a transcription more or less complete, made according to some explicit model, of the text present on those pages
- metadata about the sources which have been so treated, in particular their relationships
- metadata about the way in which the digitization and transcription have been carried out
- varying levels of annotation concerning the topics, persons, events etc. treated in the texts, their linguistic properties, etc. etc.

The TEI offers a range of ways of organizing and encoding all of these aspects
Transcription of primary sources using the TEI

- `<text>`: contains a structured reading of a document's intellectual content ... its ‘text’ (or a set of such things)
- `<facsimile>`: organizes a set of page (vel sim) images representing a document
- `<sourceDoc>`: a structured representation of a document considered purely as a physical object, an ‘objective’ transcription
- `<teiHeader>`: provides metadata describing the objects concerned, the encoding and analytic methods applied, notably including a `<msDesc>`

A `<TEI>` element contains at least a `<teiHeader>`, followed by as many of the others as you wish to include.
In the simplest case, we just want to organize a series of page image files so that an application will display them correctly.

```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- metadata concerning our edition -->
  </teiHeader>
  <facsimile>
    <graphic url="page1r.png"/>
    <graphic url="page1v.png"/>
    <graphic url="page2r.png"/>
    <graphic url="page2v.png"/>
  </facsimile>
</TEI>
```

This method lacks structure...
A slight improvement

(Or we could use `<sourceDoc>` in place of or as well as `<text>` depending on our editorial principles)
However....

- where do we show that these are alternate images of the same page?
- where do we show that these pages are linked in some way (for example, as a leaf, or a gathering)?
- where do we record metadata about the images themselves?
Support for multiple images of the same surface?

The `<surface>` element allows us to group alternative images:

```xml
<facsimile>
  <graphic url="page1.png"/>
  <surface>
    <graphic url="page2-highRes.png"/>
    <graphic url="page2-lowRes.png"/>
  </surface>
  <graphic url="page3.png"/>
  <graphic url="page4.png"/>
</facsimile>
```
And surface grouping?

The `<surfaceGrp>` element allows us to group surfaces:

```xml
<facsimile>
  <surfaceGrp type="leaf">
    <surface>
      <graphic url="page1recto.png"/>
    </surface>
    <surface>
      <graphic url="page1verso.png"/>
    </surface>
  </surfaceGrp>
</facsimile>
```
Sub-parts of a surface?
The <zone> element allows us to indicate any region within a surface

- A <zone> identifies a polygon (not necessarily rectangular): any two-dimensional space
- It is defined either using the @points attribute (borrowed from SVG) or using the attributes @ulx, @uly, @lrx and @lry (borrowed from xhtml)
- The points defining a zone must use the coordinate system defined for the surface
- A coordinate system defines a range of values for the (x,y) point-pairs defining a two-dimensional polygon: not a measurement

```xml
<facsimile>
  <surface ulx="0" uly="0" lrx="40" lry="30">
    <graphic url="page1r.png"/>
    <zone points="22,10 30,21 17,25 12,23">
      <graphic url="page1rdetail.png"/>
    </zone>
  </surface>
</facsimile>
```
Aligning images and transcription

- The @facs attribute, available on any transcriptional element, points to a <zone>, <surface>, or (in the simplest case) a <graphic>

- (And the @start attribute on <zone> or <surface> can also point into a transcription)

```xml
<facsimile>
  <surfaceGrp type="leaf">
    <surface xml:id="p1r">
      <graphic url="page1r.png"/>
      <graphic url="page1r.tiff"/>
    </surface>
    <surface xml:id="p1v">
      <graphic url="page1v.png"/>
    </surface>
  </surfaceGrp>
</facsimile>

<text>
  <pb facs="#p1r"/>
  <!-- text from page 1 recto transcribed here -->
  <pb facs="#p1v"/>
  <!-- text from page 1 verso transcribed here -->
</text>
```
For example (1)

We can identify several distinct zones in this page:

- The heading
- the ornamental capital
- the picture of a bell

... They are all rectangular zones so we can identify them just by specifying their opposing corners.
de Géométrie.

D V S O N E T A C C O R D
D E S C L O C H E S , E T D E S A L-
L E U R E S D E S C H E U X A U S , C H A I R O N S & C H A R G E S:
L A D I M E N T D U C O R P S H U M A N .

Chapitre Septième.

Les cloches ont quasi des figures maintenant des ronde pyramides impar-
aires & irrégulières: & leur axe, se fait par règle Geo-
metrique: comme si les deux cloches (qui sont pendantes à un même axe, ou en y-
montrent bicy du que leur accord se fer-
rà en contraires parti-
ties, comme vous voyez ici figuré: car
quand l'une fera en haut, l'autre decline-
ra en bas. Autrement
si elles de-
De Geometrie 159

DU SON ET ACCORD DES CLOCHES, ET DES ALEURES DES CHEVAUX, CHARIOTS & CHARGES: DES FONTAINES, & ENCYCLIE DU MONDE: & DE LA DIMENSION DU CORPS HUMAIN.

Chapitre Septiesme.

Le son & accord des cloches pendans en un mesme axe, est fait en contraires parties.

Es cloches sont quasi figures de ronnes pyramides imperfaictes & irregulieres: & leur accord se fait par reigle Geometrique: comme si les deux cloches C & D sont pendantes a vn mesme axe, ou essieu, A B: je dy que leur accord se fera en contraires parties. Autremet si elles de...
Surfaces and zones...

The relationship between surface and zone can be quite complex
Multi-part surfaces

Zones can cross surface and zone boundaries:

More of this later!
Transcription : a can of worms

What's going on here?

1. ‘agreeable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreeable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreeable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.
Transcription: a can of worms

What's going on here?

1. ‘agreeable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreeable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreeable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.
Transcription: a can of worms

What's going on here?

1. ‘agreeable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.

2. ‘agreeable’ is deleted and replaced by ‘pleasing’

3. Originally, the text read ‘agreeable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.
Transcription: a can of worms

What's going on here?

1. ‘agreeable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreeable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreeable’, but at some subsequent stage this word was deleted; the word ‘pleasing’ was added in the same context.
Transcription: a special kind of reading

What is the goal of a transcription?

- to make a primary source accessible ...
- ... and comprehensible
- which may imply adding or using much additional information

Hence,

- all transcription is selective
- all transcription is imaginative

TEI distinguishes between documentary and textual transcription
Making a documentary transcription

The `<sourceDoc>` element allows us to represent ‘uninterpreted’ text within a document

- The `<sourceDoc>` element contains `<surface>` and `<zone>` elements, just like a `<facsimile>` ...
- ... except that its components contain transcribed text as well as (or instead of) images
- A special kind of zone, called `<line>` is also available
- along with a small number of neutral tags for obvious metatextual interventions
Layers of transcription

- The palaeographic layer: what characters do we see here?
- The documentary or diplomatic layer: what was actually written on the page?
- The editorial or semantic layer: how should it be read?
Palaeographic layer

- identify the marks we consider to be letters
- map the letters to an appropriate unicode character
- decide which non-standard or variant characters we need to preserve

The TEI <g> element is your friend...
& demy trebuchans de
soixante sept pieces au marc
a ung felin & demy de remede
Cours chacune piece po\textordfeminine;
cinqu\textordertrees sousz tourois.
Sur le pris de quinze livres
tourois le marc dargent le roy
fust c\textordertree la fabricaci\textordertrees des testons a dix deniers
dix huict grains troys quartz
de fin qui valt unze deniers
six grains dargent le roy, a
In a textual transcription, by contrast ...

We use traditional TEI structuring elements (<div>, <head>, <p> etc.)

We make explicit a range of interventions in the text, such as:

- original layout information
- abbreviations or other arcana
- ‘evident’ errors which invite correction or conjecture
- scribal additions, deletions, substitutions, restorations
- non-standard orthography (etc.) which invites normalisation
- irrelevant or non-transcribable material
- passages which are damaged or illegible

More of all this later!
Within `<text>` the logical view is privileged, but the physical view can ‘show through’ as empty milestone elements:

- `<gb>` the start of a new gathering or quire
- `<pb>` the start of a new page
- `<cb>` the start of a new column
- `<lb>` the start of a new written line

These are primarily useful to establish a reference system.

The `<fw>` element can be used to mark ‘paratextual’ features such as running heads, foliotation etc.

The `<handShift>` element can be used to mark changes of hand or writing in a document.
<p>
<!-- ... -->

<pb n="5r"/>
<fw place="topRight" type="pageNum">5</fw>

<lb/>

<expan>et</expan> demi trebuchans de <lb>soixante sept pieces au marc <lb>a ung felin <expan>et</expan> demi de remede <lb>Cours chacune piece <expan>pour</expan> <lb></expan>

<lb/>

<expan>cinquante</expan> soubz <expan>tournois</expan> <pc></pc>
</p>

<p>
<br/>

<lb>Sur le pris de quinze livres <lb>le marc dargent le roy <lb>fust <expan>continuee</expan> la <expan>fabricacion</expan> <lb>des testons a dix deniers <lb>dix huit grains troys quartz <lb>de fin qui valent <expan>unze deniers <lb>six grains dargent le roy, a</expan> 
</p>
Editorial strategy may be simply to note that we have expanded the abbreviations:

```html
<p>
  <lb/>Cours chacune piece <expan>pour</expan>
  <lb/>
  <expan>cinquante</expan> soubz <expan>tournois</expan>
  <pc>.
</p>
```
Abbreviation example (2)

As you noticed, ‘pour’ was actually written ‘po’ followed by an ‘r’ superscript; ‘cinquante’ as ‘cinquāte’ with a macron on the ‘a’ to indicate nasalisation.
A more diplomatic encoding might therefore be:

```xml
<p>
  <abbr>po&#xFFFD;</abbr> .... <abbr>cinquāte</abbr>
</p>
```

And of course TEI permits both cake and the eating off it:

```xml
<choice>
  <abbr>po<am>&#xFFFD;</am></abbr>
  <expan>po<ex>u</ex>r</expan>
</choice>
```
Normalisation example

In this case, a further semantic regularisation is possible:

```xml
<lb/>
dix <choice>
   <orig>huict</orig>
   <reg>huit</reg>
</choice> grains
<choice>
   <orig>troys quartz</orig>
   <reg>trois-quart</reg>
</choice>
```

```xml
<lb/>
<measure quantity="18.75" unit="gr">dix
   <choice>
      <orig>huict</orig>
      <reg>huit</reg>
   </choice> grains <choice>
      <orig>troys quartz</orig>
      <reg>trois-quart</reg>
   </choice>
</measure>
```
Some difficulties

These methods are perfectly adequate in simple cases. They rapidly encounter problems when:

- overlap happens (as it always does)
- the sequence of interventions is important or indeterminate
- the layout and the meaning of the writing are not easily separable

Work-arounds do exist for all of these, of course
Overlap happens

<line>exigences de la philosophie. Dans la

<citation>"Contribution à la critique" de la philosophie du droit de Hegel", il se moque ...

</citation>
Public/private
DONNE DES COULEURS CLAIRES
CAR DEPART
JE SUIS L'HEURE
QUI ON ME ARRIVE
MALGRE CEUX PLEURANT
GAIEMENT
MAJESTEMENT NOUS SERONS TOUS JAMAIS DES ONUS

QUI DEMEURE EST DE VOUS PARTIELLEMENT
destinee et de vous reviendrons que sont
parties
Je te vois jamais personne, jamais voiles
sacre aujourd'hui, depuis mes
Our interest may be to trace the development or evolution of a text, as witnessed by various documents.

But the sense of a text may be inseparable from the document in which it is manifest, because:

- the sense is partly or entirely carried by its visualisation
- the document is constructed in a non-linear or combinatory manner, with the explicit goal of generating multiple meanings, many ‘texts’

The goal of a TEI encoding remains to make explicit one or several views on a set of documents, to facilitate an analysis of the way these documents (and the readings of them) interact.
One may contemplate, with equanimity, every complexity of Byzantine medieval military history but be quite defeated by the unfamiliar vocabulary of the mysteriously interconnected universe which is the TEI.


‘almost without exception, no scholarly electronic edition has presented material which could not have been presented in book form ... most electronic scholarly editions [fail] to use new computer methodologies to explore the texts which they present’

http://computerphilologie.uni-muenchen.de/jg03/robinson.html (2005)
So what should we anticipate doing with the editions of the future?

- Better visualisation tools
- Better access mechanisms, for both metadata and data
- Tools for dynamic hypothesis testing
- More use of social networking and crowd sourcing
New technical paradigms

Visualisation and analysis at two levels

- sub-documentary components
- across corpora of documents
- locating and presenting for patterns of variation

quantitative codicology meets evolutionary biology
The wisdom of crowds and the demise of the editor

‘We are all engaged in the business of understanding: distributed editions fashioned collaboratively may become the ground of our mutual enterprise.’ (PMWR, 2007)

- Transcribe Bentham: http://www.ucl.ac.uk/transcribe-bentham/
- Oxyrhynchus papyri: ancientlives.org

‘What is needed is a commitment to cooperative work among developers in a chaotic environment of experimentation and communication.’ (CMSMCQ, 1996)