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 digutization 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.

A digital facsimile edition

In the simplest case, we just want to organize a series of page image files so that an application will display them correctly.

```
<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

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
<!-- metadata concerning our edition-->
  </teiHeader>
  <text>
    <pb  facs="page1.png"/>
<!-- transcript of page 1 here -->
    <pb  facs="page2.png"/>
<!-- transcript of page 2 here -->
  </text>
</TEI>
```

(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 altermative images :

```
<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:

```
<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* attrib ute (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

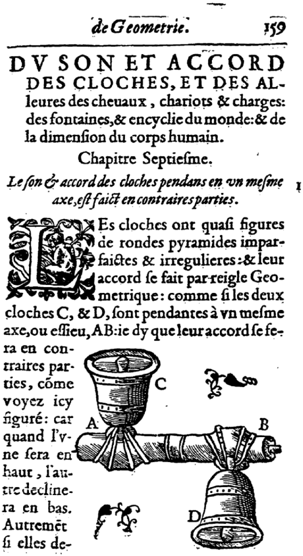
```
<facsimile>
  <surface  ulx="0"  uly="0"  lrx="40"
    lry="30">
    <graphic  url="page1r.png"/>
    <zone  points="22,10 30,21 17,25 12,23">
      <graphic  url="page1rdetail.png"/>
    </zone>
  </surface>
</facsimile>
```

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)

```
<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)



de Geometrie. 159

DV SON ET ACCORD DES CLOCHES, ET DES ALLEURES DES CHEUAUX, chariots & charges: des fontaines, & encyclie du monde: & de la dimenſion du corps humain.

Chapitre Septieſme.

Le ſon & accord des cloches pendans en vn meſme axe, eſt faict en contraires parties. 1

LEs cloches ont quaſi figures de rondes pyramides imparfaictes & irregulieres: & leur accord ſe fait par reigle Geometrique: comme ſi les deux cloches C, & D, ſont pendantes à vn meſme axe, ou eſſieu, AB: ie dy que leur accord ſe fera en contraires parties, cõme voyez icy figuré: car quand l'vne ſera en haut, l'autre declinera en bas. Autremẽt ſi elles de-

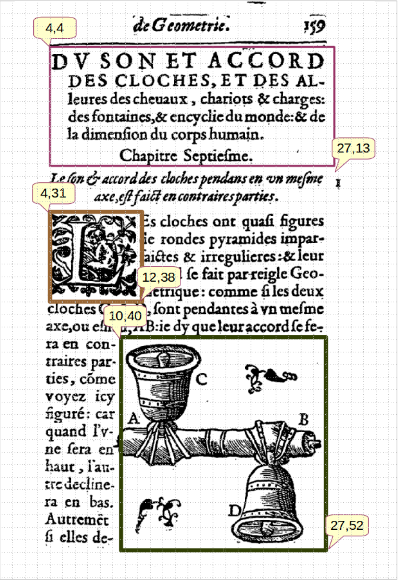
We can identify several distinct zones in this page:

- 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

... like this :



de Geometrie. 159

DV SON ET ACCORD DES CLOCHES, ET DES ALLEURES des cheuaux, chariots & charges: des fontaines,& encyclie du monde:& de la dimenſion du corps humain.

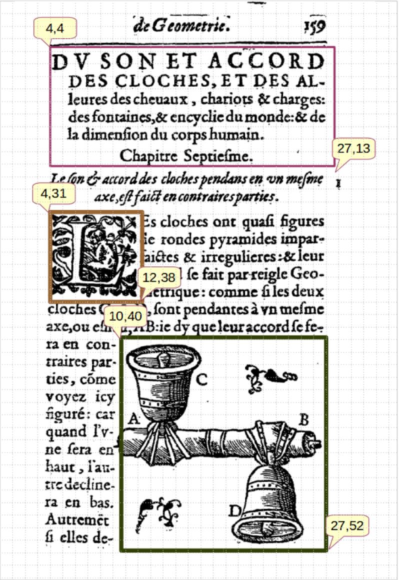
Chapitre Septieſme.

Le ſon & accord des cloches pendans en vn meſme axe,eſt faict en contraires parties. 1

Es cloches ont quaſi figures ie rondes pyramides impar aictes & irregulieres:& leur ſe fait par reigle Geo etrique: comme ſi les deux cloches C ,ſont pendantes à vn meſme axe,ou e ,AB:ie dy que leur accord ſe fera en contraires parties, cõme voyez icy figuré: car quand l'vne ſera en haut, l'autre declinera en bas. Autremẽt ſi elles de-

```
<facsimile>
  <surface ulx="0" uly="0" lrx="52"
    lry="32">
    <graphic url="bovelles.png"/>
    <zone ulx="4" uly="4" lrx="27"
      lry="13"/>
<!-- the title -->
    <zone ulx="4" uly="31" lrx="12"
      lry="38"/>
<!-- the capital -->
    <zone ulx="10" uly="40" lrx="27"
      lry="52"/>
<!-- the bell -->
  </surface>
</facsimile>
```

And the transcription



de Geometrie. 159

DV SON ET ACCORD
DES CLOCHES, ET DES AL-
leures des cheuaux , chariots & charges:
des fontaines,& encyclie du monde:& de
la dimenſion du corps humain.
Chapitre Septieſme.

*Le ſon & accord des cloches pendans en vn meſme
axe,eſt faict en contraires parties.*

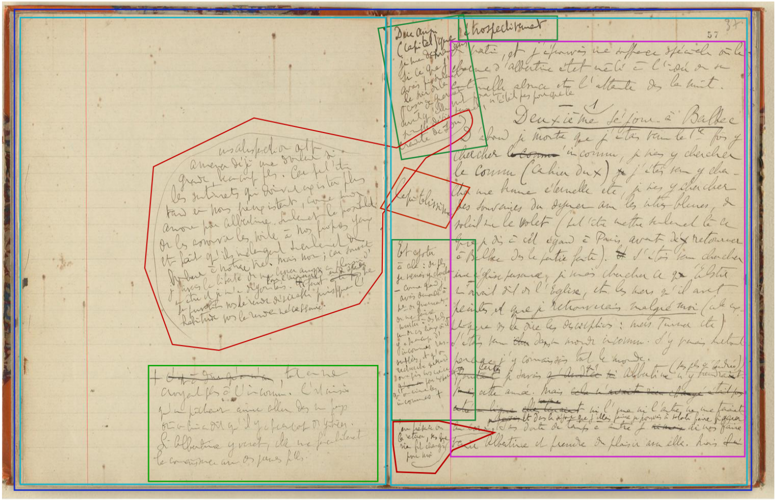
1

L Es cloches ont quaſi figures
ie rondes pyramides impar-
aictes & irregulieres:& leur
l ſe fait par reigle Geo-
etrique: comme ſi les deux
cloches C [illegible], ſont pendantes à vn meſme
axe,ou eſſ[illegible]u, AB:ie dy que leur accord ſe fe-
ra en con-
traires par-
ties, cõme
voyez icy
figuré: car
quand l'v-
ne ſera en
haut, l'au-
tre decline-
ra en bas.
Autremẽt
ſi elles de-

```
<facsimile>
  <surface  xml:id="B49r"  ulx="0"  uly="0"
    lrx="52"  lry="32">
    <graphic  url="bovelles.png"/>
    <zone  xml:id="B49rHead"  ulx="4"
      uly="4"  lrx="27"  lry="13"/>
<!--the title -->
    <zone  xml:id="B49rCap"  ulx="4"
      uly="31"  lrx="12"  lry="38"/>
<!-- the capital -->
    <zone  xml:id="B49rFig"  ulx="10"
      uly="40"  lrx="27"  lry="52"/>
<!-- the bell -->
  </surface>
</facsimile>
<text>
  <body>
    <pb  facs="#B49r"/>
    <fw>De Geometrie 159</fw>
    <head  facs="#B49rHead"> DU SON ET ACCORD DES
CLOCHES ET DES ALleures des cheuaux, chariots &amp;
charges: des fontaines,&amp; encyclie du monde: &amp;
de la dimension du corps humain.</head>
    <head>Chapitre Septiesme.</head>
    <div  n="1">
      <p>Le son &amp; accord des cloches pendans en
ung mesme axe, est faict en contraires parties.</p>
      <p>
        <g  facs="#B49rCap">L</g>Es cloches ont quasi
figures de rondes pyramides imperfaictes &amp;
irregulieres: &amp; leur accord se fait par reigle
Geometrique: comme si les deux cloches C &amp; D sont
pendantes à vn mesme axe, ou essieu, A B: je dy que
leur accord se fera en contraires parties
co<ex>m</ex>me voyez icy figuré. Car quand l'vne sera
en haut, l'autre declinera en bas. Autreme<ex>m</ex>t
si elles de.<figure  facs="#B49rFig1"/>
```

Surfaces and zones...

The relationship between surface and zone can be quite complex

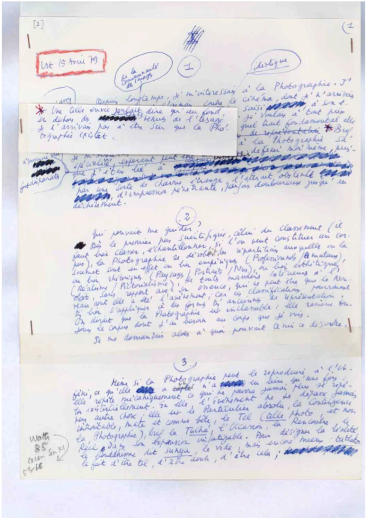


:

Source gallica.bnf.fr / Bibliothèque nationale de France

Multi-part surfaces

Zones can cross surface and zone boundaries :



[2] (1

Lnt 15 Avril 79

1

de la communauté des images

idiotique

ans

Depuis longtemps, je m'intéressais à la Photographie. J'[illegible] contre le cinéma, dont je n'arrivais [illegible] saisi [illegible] à son égard. Je voulais à tout prix [illegible] quel trait fondamental elle [illegible] ~~de représentation~~ * Bref [illegible] à la Photographie sa [illegible] défendue ma même, pas [illegible] spécialité, espérant peut-être [illegible] que j'étais lié à [illegible] par une sorte de charme étrange d'attrait obstiné [illegible], d'impression pénétrante, parfois douloureuse jusqu'au déchirement.

* Une telle envie voulait dire qu'au fond, en dehors des évidences venues de l'usage je n'arrivais pas à être sûr que la Photographie existât.

n'y prenne que [illegible] fort disparates

2

Qui pouvait me guider? Dès le premier pas scientifique, celui du classement (il faut bien classer, échantillonner, si l'on veut constituer un corpus), la Photographie se dérobait. Les répartitions auxquelles on la soumet sont en effet ou bien empiriques (Professionnels / Amateurs), ou bien rhétoriques (Paysages / Portraits / Nus), ou bien esthétiques (Réalisme / Pictorialisme), de toutes manières extérieures à l'objet, sans rapport avec son essence, qui ne peut être que le Nouveau dont elle a été l'avènement ; car ces classifications pourraient très bien s'appliquer à des formes très anciennes de représentation. On dirait que la Photographie est inclassable : elle ramène toujours le corpus dont j'ai besoin au corps que je vois.

Je me demandais alors à qui pouvait la lui désœuvre.

3

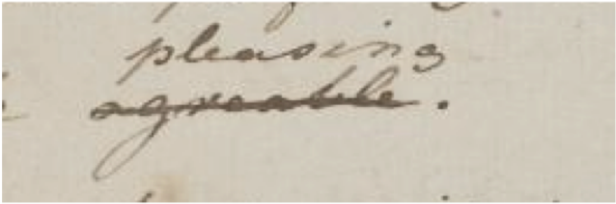
Même si la Photographie peut le reproduire à l'infini, ce qu'elle a ~~capté~~ vu n'a [illegible] eu lieu qu'une fois : elle répète mécaniquement ce qui ne pourra jamais plus se répéter existentiellement. En elle, l'Evènement ne se dépasse jamais vers autre chose ; elle est le Particulier absolu, la Contingence inimitable, mate et comme bête, le Tel (telle photo, et non la Photographie), bref la Tuché, l'Occasion, la Rencontre, le Réel, dans son expression infatigable. Pour désigner la réalité, le bouddhisme dit sunya, le vide, mais encore mieux : tathata, le fait d'être tel, d'être ainsi, d'être cela ;

Watts 85 Sun XI [illegible] 53 66

More of this later !

Transcription : a can of worms

What's going on here?

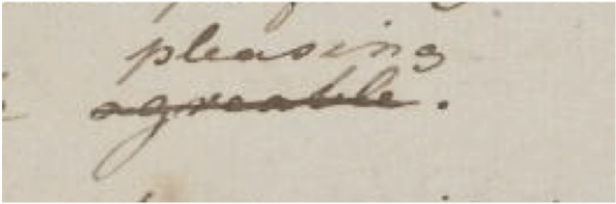


pleasing
~~agreable~~.

1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, 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?



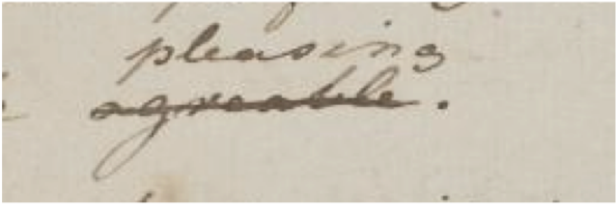
pleasing

~~agreable~~.

1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, 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?

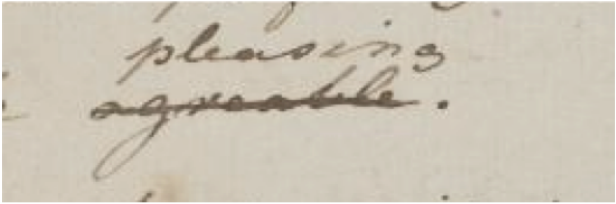


pleasing
~~agreable~~.

1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, 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?



pleasing

~~agreable~~.

1. ‘agreable’ is struck-through, ‘pleasing’ is written above it, in the interlinear space.
2. ‘agreable’ is deleted and replaced by ‘pleasing’
3. Originally, the text read ‘agreable’, 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

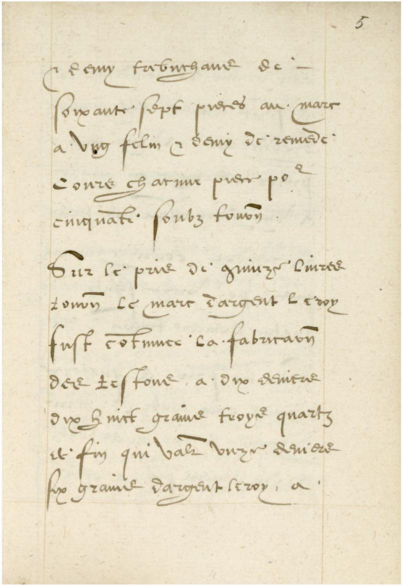
TEI distinguishes between documentary and textual transcription

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



5

et demy forburgance de —
soixante sept pieces au marc
a vng felin et demy de remede
Courre chascune piece pour
cinquante solz tournois.
Sur le pris de quinze livres
tournois le marc d'argent le roy
fust commencee la fabricacion
desd. testons, a dix deniers
dix huit grains troys quartz
de fin qui vault unze deniers
six grains d'argent le roy, a

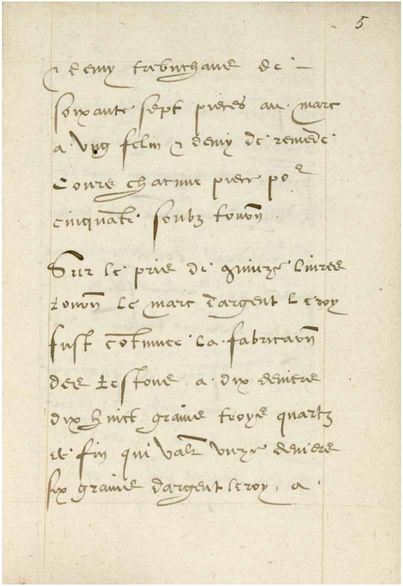
- The palaeographic layer : what characters do we see here?
- The documentary or diplomatic layer : what was actually written on the page?
- The editorial or semantic layer : how should it be read?

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...

Documentary transcription of page 5



```
<surface  n="5r">
  <zone>5</zone>
  <line>& demy trebuchans de</line>
  <line>soixante sept pieces au marc</line>
  <line>a ung felin & demy de remede</line>
  <line>Cours chacune piece po&#xFFFD;</line>
  <line>cinquāte soubz tourois.</line>
  <line>Sur le pris de quinze livres </line>
  <line>tourois le marc dargent le roy </line>
  <line>fust cõtinuee la fabricaciõ</line>
  <line>des testons a dix deniers </line>
  <line>dix huict grains troys quartz</line>
  <line>de fin qui valt unze deniers </line>
  <line>six grains dargent le roy, a</line>
</surface>
```

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!

Original layout information

Within <text> the logical view is privileged, but the physical view can ‘show through’ as empty milestone elements :

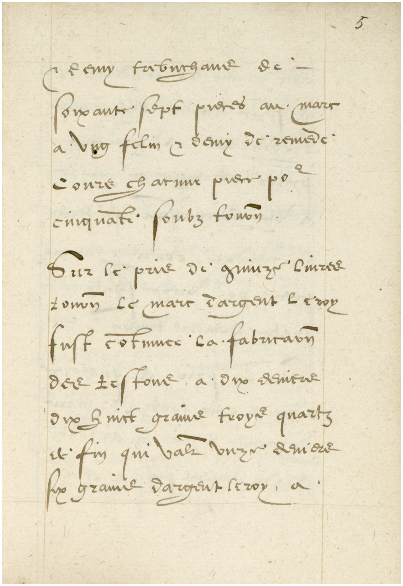
- <gb> the start of a new gathering or quire
- <pb> the start of a new page
- <cb> the start of a new column
- <lb> the start of a new written line

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.

Textual transcription of page 5



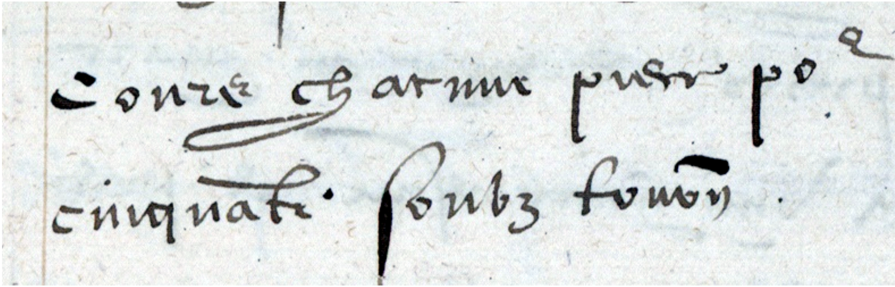
5

et demy trebuchans de
soixante sept pieces au marc
a ung felin et demy de remede
Cours chacune piece pour
cinquante soubz tournois.

Sur le pris de quinze livres
tournois le marc dargent le roy
fust continuee la fabricacion
des testons a dix deniers
dix huict grains troys quartz
de fin qui valent unze deniers
six grains dargent le roy, a

```
<p>
<!-- ... -->
  <pb n="5r"/>
  <fw place="topRight" type="pageNum">5</fw>
  <lb/>
  <expan>et</expan> demy trebuchans de <lb/>soixante
sept pieces au marc <lb/>a ung
felin <expan>et</expan> demy de remede <lb/>Cours
chacune piece <expan>pour</expan>
  <lb/>
  <expan>cinquante</expan> soubz
<expan>tournois</expan>
  <pc>.</pc>
</p>
<p>
  <lb/>Sur le pris de quinze livres <lb/>
  <expan>tournois</expan> le marc dargent le roy
<lb/>fust <expan>continuee</expan> la
<expan>fabricacion</expan>
  <lb/>des testons a dix deniers <lb/>dix huict grains
troys quartz <lb/>de fin qui
<expan>valent</expan> unze deniers <lb/>six grains
dargent le roy, a
<!-- ... -->
</p>
```

Abbreviation example (1)



Editorial strategy may be simply to note that we have expanded the abbreviations:

```
<p>
 <lb/>Cours chacune piece <expan>pour</expan>
 <lb/>
 <expan>cinquante</expan> soubz <expan>tournois</expan>
 <pc>.</pc>
</p>
```

Abbreviation example (2)

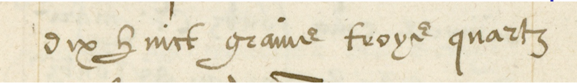
As you noticed, ‘pour’ was actually written ‘po’ followed by an ‘r’ superscript; ‘cinquante’ as ‘cinquāte’ with a macron on the ‘a’ to indicate nasalisation.
A more diplomatic encoding might therefore be :

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

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

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

Normalisation example



dix huict grains troys quartz

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

In this case, a further semantic regularisation is possible :

```
<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>
```

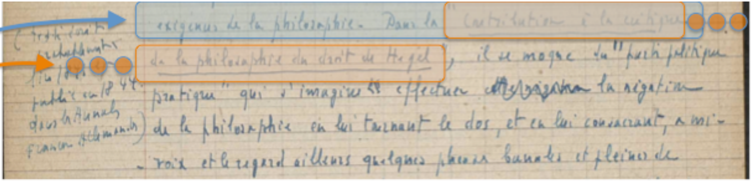
Some difficulties

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

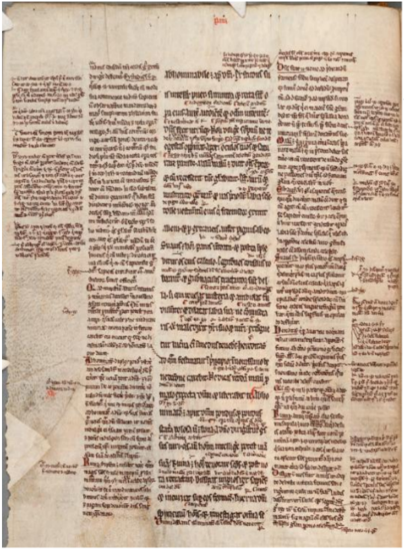
Overlap happens

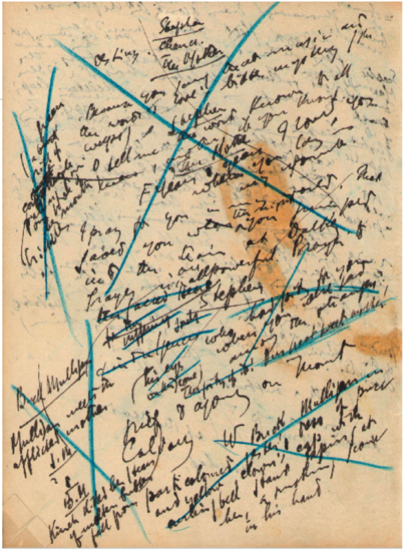


(Écrit fin 1843 publié en 1844 dans les Annales franco-allemandes) exigences de la philosophie. Dans la "Contribution à la critique de la philosophie du droit de Hegel", il se moque du "parti politique pratique" qui s'imagine effectuer ~~[illegible]~~ la négation de la philosophie en lui tournant le dos, et en lui consacrant, à mi-voix et le regard ailleurs quelques phrases banales et pleines de

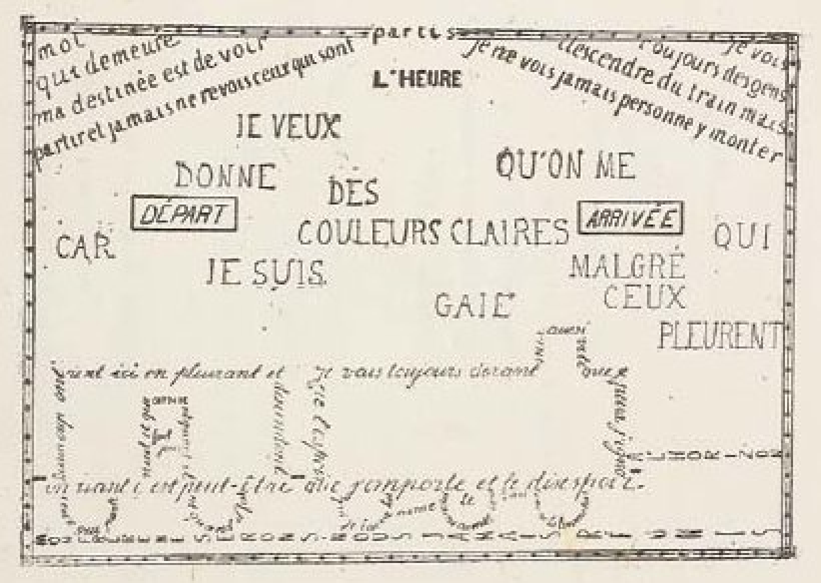
<line>exigences de la philosophie. Dans la <citation>*"Contribution à la critique*</line> de la philosophie du droit de Hegel"</citation>, il se moque ...

Public/private





Text/ image


moi qui demeure ma destinée est de voir partir et jamais ne revois ceux qui sont
partis
je ne vois jamais personne y monter je vois toujours des gens descendre du train mais
L'HEURE
JE VEUX
DONNE
DES
QU'ON ME
DÉPART
ARRIVÉE
CAR
COULEURS CLAIRES
QUI
JE SUIS
MALGRÉ
GAIE
CEUX
PLEURENT
vient ici en pleurant et je vais toujours devant
en riant c'est peut-être que j'emporte et le désespoir

Text in flux

- 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

Robinsonian provocations

> 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.

`http://www.digitalmedievalist.org/journal/1.1/robinson/#robinson.dm.1.1.0140`(2005)

> ‘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/`
- Oxyrnchus papyri : `ancientlives.org`

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