Transcribing a manuscript with TEI

What does ‘digitization’ mean?

(Not the same as ‘digitalization’!)

- production of digital images of the pages of a manuscript – a *facsimile*
- production of a *transcription* of the content of a manuscript

For manuscripts, the two are often complementary

Facsimile and transcription together

Был тогда день субботний, и, когда солнце взошло, сошлись оба полка. Немцы же и чудь пробились свиньей сквозь полки, и была тут сеча зла и велика с немцами и чудью; и стоял треск от ломающихся копий и звон от ударов мечей; и казалось, что озеро замерзшее двинулось, и не было видно льда, ибо все было залито кровью.

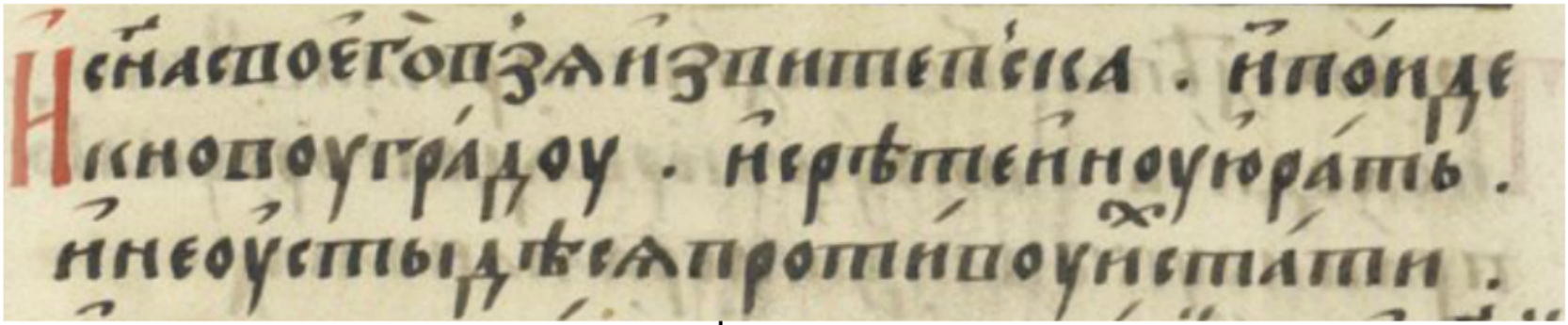
Бе тогда день суботный, солнцу восходящю и съступишася окон пълцы. Немцы же и чюдь прокишася свинъскы сквозе полки, и кысть ту сеча зла и велика немцем и чюди, и трус, от коней ломление и звук от мечнаго сечения, якоже морю помъръшю двигнутися, и не ке видети леду, покрыло ко есть все кровью.

И я слышал от очевидца, рассказавшего мне, что видел полки Божьи в воздухе, пришедшие на помощь великому князю Александру. И победил он с Божьею помощью, святой Софии и святых мучеников Бориса и Глеба, которые за нее пролили кровь свою. И побежали враги, и рубили их, гоня, как по воздуху, и некуда им было скрыться; и били их на льду на протяжении семи верст до Суболичского берега. И было убито 500 немцев, а чудя – бесчисленное множество. В плен взяли 50 знатных мужей, храбрых воевод, и привели их в Новгород, а иные в воде утонули, а иные жестоко ранены были и убежали. Был же бой этот 5 апреля в день Похвалы святой Богородицы, на память святого мученика Клавдиана. Здесь же прославил Бог великого Александра перед всеми полками, как Иисуса Навина у Иерихона. «Вот, – говорили немцы, – возьмем великого князя Александра руками», а Бог ему их предал в руки его. И не было равного ему противника в бранях.

Се же слышах от самовидца, рече ми, яко видех пълкы Божна на воздусе, пришедше на помощь великому князю Александру, и победив Божиею силою, и святыя Софии и святою мученику Бориса и Глеба, своже ради кровь свою пролияша. И даша ратнии плещи свои, и секахуть гоняше аки по леду, и не ке им камо утещи, и кнша их на семи кръстах по леду до Суболичьскаго берега, и паде немец 500, а чюди бесчисленое множьство, а руками яша немець 50 мужь, нарочитых мужей, силных воевод, и приведоша их к Новгород, а иных вода потопи, а инии зле язвени быша, и отбегоша. Бысть же той сий апреля в 5, на похвалу святыя Богородица, на память святаго мученика Клавдиана, зде же прослави Бог великаго Александра пред всеми полки, яко Исуса Навкина у Ерихона, се же рекли немцы: «Имем великого Александра руками», и сиъ ему Бог предасть к руце его, и не обретеся противник ему в пранех никогда же.

http://www.akteon-elib.ru/LITsIeVOI-LIeTOPISNYI-SVOD-Russkaya-letopisnaya-istoriya-Kniga-6/#24

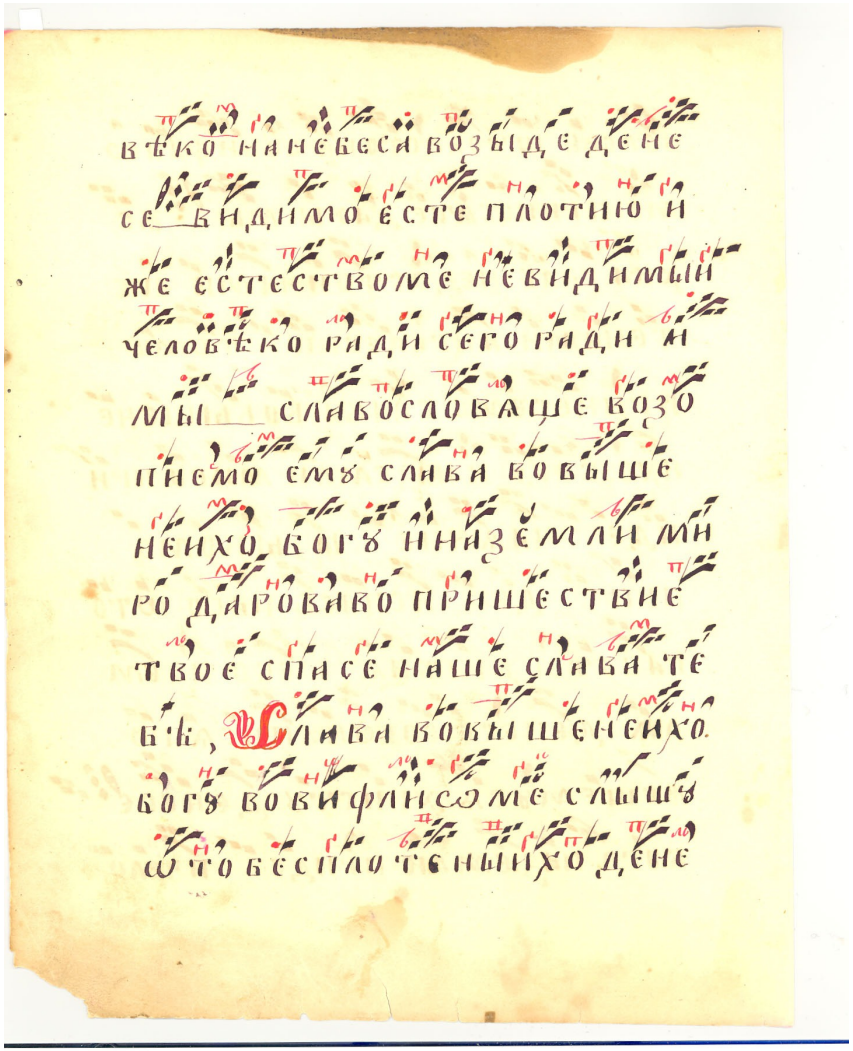
Transcription is not an automatic process



и сына своего взя из Витепска, и поиде к Новуграду, и срете иную рать, и не устыдеся противу их стати, и

а сына своего взял из Витебска и пошел к Новгороду. И встретил иную рать, и встал против них,

Transcription: a special kind of reading



MS5045 Russian Musical Manuscript, from http://faculty.goucher.edu/eng241/high_church_slavonic_leaf_images.htm

What is the goal of your transcription?

- to make a primary source accessible ...
- ... but also comprehensible
- which may imply adding (or suppressing) a lot of information

Because...

- all transcription is selective
- all transcription is imaginative

Transcription

What does a transcription add to a simple facsimile?

Transcribers typically try to make explicit :

- (some) original layout information
- abbreviations and other strange symbols
- ‘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

What kind of transcription do you want?

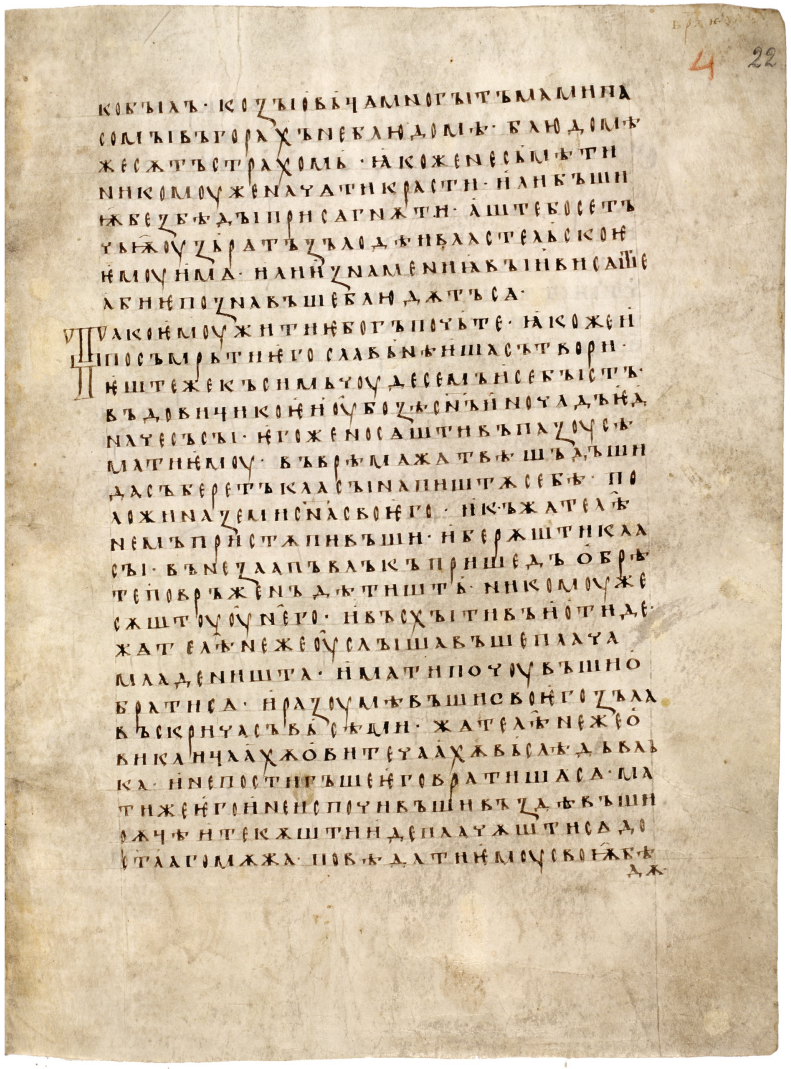
- <teiHeader>: provides metadata for the whole thing, at various levels, typically including a <msDesc>
- <text>: contains a structured reading of a document's intellectual content ... its 'text'
- <facsimile>: organizes a set of page images representing a document
- <sourceDoc>: a non-interpretative transcription of a physical document, e.g. for a *dossier génétique*

Does your transcription represent a 'text' or a 'document' ?

How is your transcription organised?

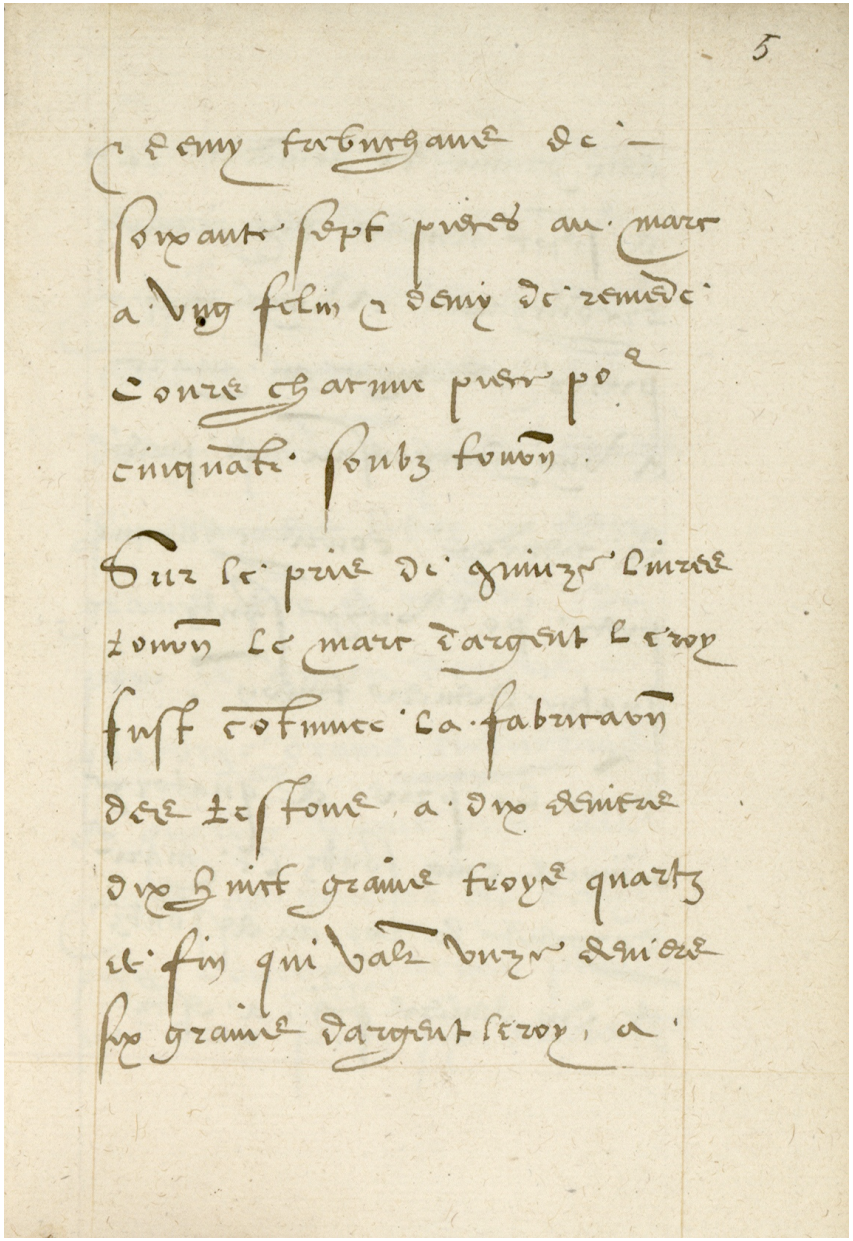
- Just pages or folios, composed of blocks or lines
- Sections, paragraphs, verse lines, lists, sentences ...
- Or both?

Layers of transcription



- Paleographic level : identification of characters and graphemes
- Documentary (or diplomatic) level: decide what has been written
- Editorial or semantic level : decide how it should be read

A typical minimal encoding

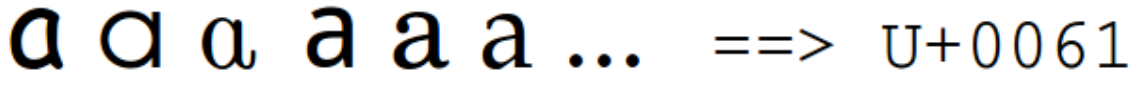
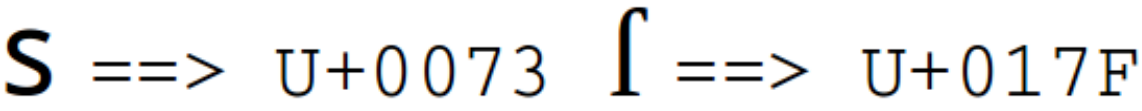


```
<pb n="13" xml:id="B452346101_Ms629_0015"/>
<fw place="top-right" type="pageNum">5</fw>
<lb/>&amp; demy trebuchans de
<lb/>soixante sept pieces au marc
<lb/>a ung felin &amp; demy de remede
<lb/>Cours chacune piece <expan>pour</expan>
<lb/><expan>cinquante</expan> soubz
     <expan>tournois</expan><pc>.</pc></p>
<lb/><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<pc>,</pc> a
```

Representation of the physical structure

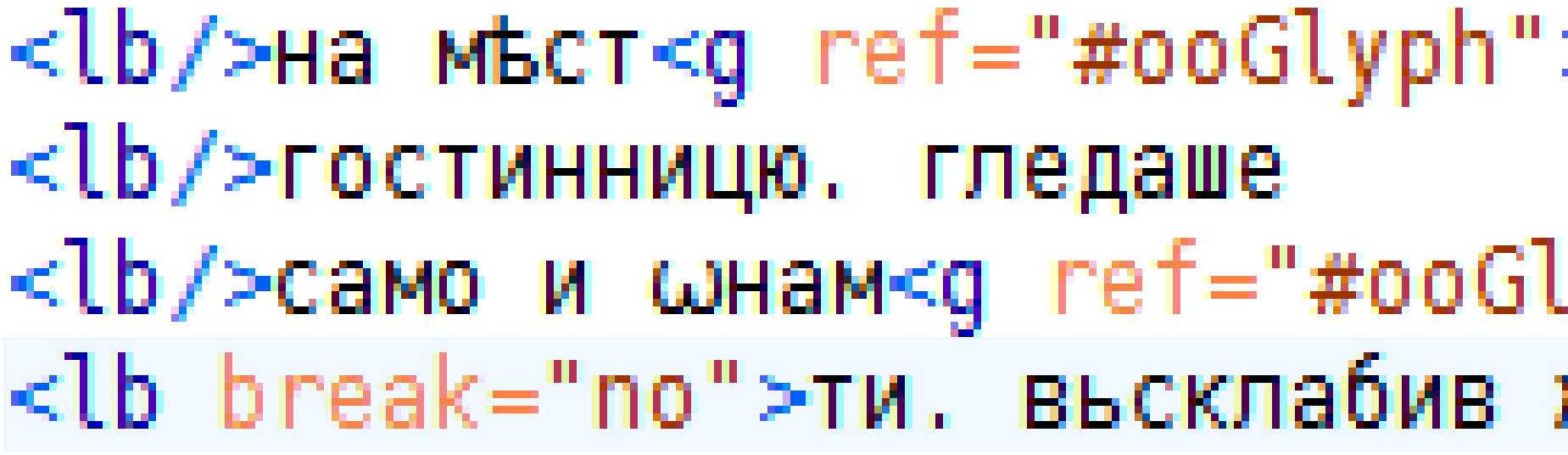
- The physical organisation of a manuscript (its binding, folios, leaves, pages, columns) rarely, if ever, corresponds with its logical organisation (sections, chapters, paragraphs, lines)
- Whichever we choose to represent in our XML structure, we will have to represent the other using empty 'milestone' elements
- For example, in the logical view, we can use <gb>, <pb>, <cb>, or <lb> to indicate the start of gatherings, pages, column or lines
- Or in the physical view, we could use a <milestone> to indicate the starts of divisions, paragraphs, etc.

Characters and glyphs

- the same character may be represented in many different forms
  - e.g. 
  - e.g. 
- the character or glyph we see may not yet exist in Unicode

The <g> element allows us to indicate the presence of a specific glyph, or a non-Unicode character

Using <g>



Bdinski, fol 7r, detail

There is no Unicode character for the ligatured oo here: we tag it as a <g>

```
<lb/>на мѣст<g   ref="#ooLig">оо</g>. и вьшедь вь
<lb/>гостинницю. гледаше
<lb/>само и ѡнам<g   ref="#ooLig">оо</g> хоте видѣ
<lb   break="no"/>ти. вьсклабив же се рече
```

#ooGlyph points to a description of the glyph, provided in the TE' header.

Abbreviations &c.

In Western MSS, we commonly distinguish :

Suspensions the first letter or letters of the word are written, generally followed by a point : for example 'e.g.' for 'exempla gratia'

Contractions both first and last letters are written, generally with some mark of abbreviation such as superscript strokes, or points : e.g. 'Mr.' for 'Mister'

Brevigraphs Special signs such as the Tironian *nota* used for 'et', the letter p with a barred tail used for 'per', the letter c with a circumflex used for 'cum' etc.

Superscripts Superscript letters (vowels or consonants) used to indicate various kinds of contraction: e.g. 'w' followed by superscript 'ch' for 'which'.

Most of the symbols needed are available in Unicode, though not necessarily in all fonts.

Abbreviation and Expansion

An abbreviation may be viewed in two different ways:

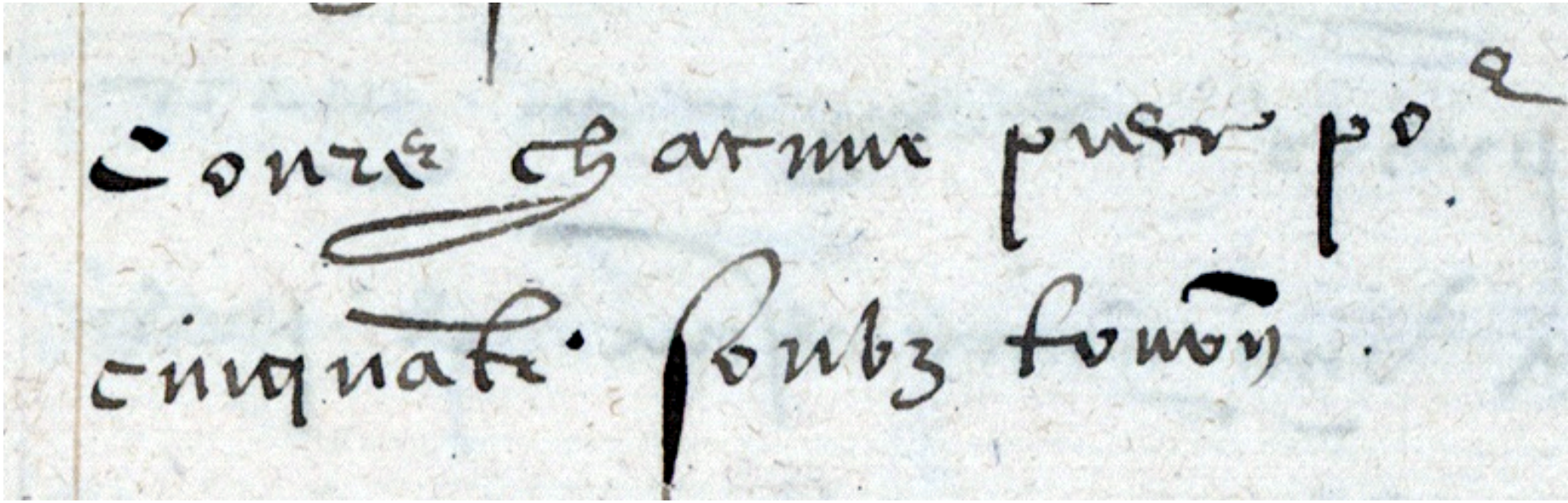
- as a particular sequence of letters or marks upon the page: thus, a ‘p with a bar through the descender’, a ‘superscript hook’, a ‘macron’
- as an alternative way of representing a sequence of letters : thus, ‘per’, ‘re’, ‘n’

Two sets of tags

TEI proposes elements for two levels of encoding:

- the whole of an abbreviated word and the whole of its expansion may be marked using <abbr> and <expan> respectively
- abbreviatory signs or characters and the 'invisible' characters they imply may be marked using <am> and <ex> respectively

A French example



We might just 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>
```

... or we might just record the abbreviated forms

As you noticed, ‘pour’ was actually written ‘po’ followed by an ‘r’ subscript; ‘cinquante’ as ‘cinquāte’ with a macron on the ‘a’ to indicate nasalisation.

```
<p>
  <abbr>po&#xFFFD;</abbr> .... <abbr>cinqu&#x0101;te</abbr>
</p>
```

... or we might look a bit closer

.We can tag the abbreviation markers and the expansion directly

```
<p> po<am>&#xFFFD;</am> ... or po<ex>u</ex>r </p>
```

... or within the <abbr> or <expan> as appropriate

```
<abbr>po<am>&#xFFFD;</am>
</abbr>
```

```
<expan>po<ex>u</ex>r</expan>
```

Or we might want to show there's a <choice> ...

The <choice> element wraps alternative mutually exclusive ways of encoding the same phenomenon:

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

Not intended for use with textual variants (for which, use <app>)

Types of abbreviation

The *@type* attribute on <abbr> is a useful way of categorising abbreviations, whether for statistical purposes, or to allow for different types to be rendered differently:

```
<choice>
  <abbr   type="brevigraphe">po<am>&#xFFFD;</am>
  </abbr>
  <expan   resp="#LB">po<ex>u</ex>r</expan> en <choice>
    <abbr   type="suspension">fin<am>.</am>
    </abbr>
    <expan>fin<ex>ir</ex>
    </expan>
  </choice>
</choice>
```

As elsewhere, the *@resp* and *@cert* attributes can be used to indicate who is responsible for an expansion, and the degree of certainty attached to it.

This encoding might be displayed as : ‘po(u)r en fin*ir* [LB]’

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>relea</sic>
```

```
<corr>relicta</corr>
```

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

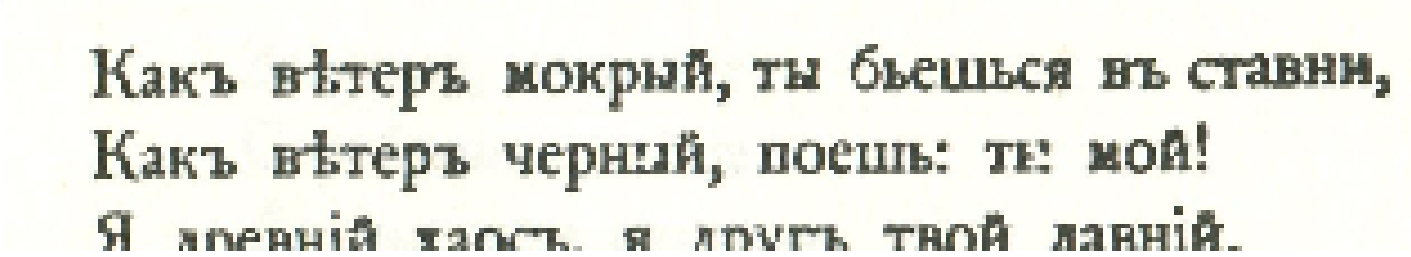
```
<choice>
  <sic>relea</sic>
  <corr   cert="high">relicta</corr>
  <corr   cert="low">relatio</corr>
</choice>
```

Normalization

Source texts rarely use modern orthography. For retrieval and other processing reasons, however, the modernized form may be needed.

The <reg> (regularized) element is available used to mark a normalized form; the <orig> (original) element to indicate a non-standard spelling. These elements can of course be grouped as alternatives using the <choice> element

A Russian example

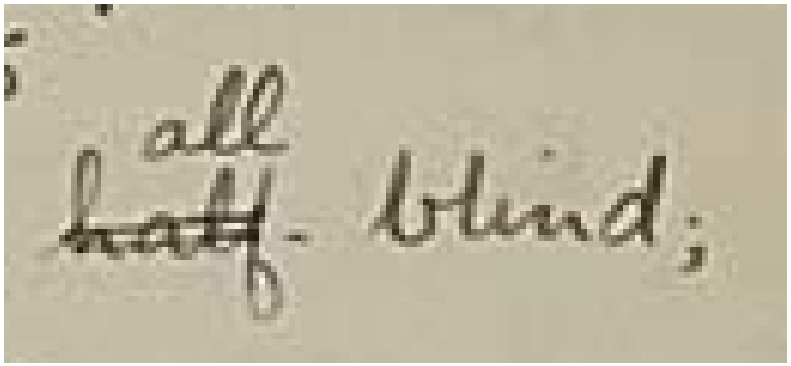

Какъ вѣтеръ мокрый, ты бьешься въ ставни,
Какъ вѣтеръ черный, поешь: ты мой!

```
<l>
  <choice>
    <orig>Какъ</orig>
    <reg>Как</reg>
  </choice>
  <choice>
    <orig>вѣтеръ</orig>
    <reg>ветер</reg>
  </choice> мокрый, ты бьешься <choice>
    <orig>въ</orig>
    <reg>в</reg>
  </choice> ставни,
</l>
<l>
  <choice>
    <orig>Какъ</orig>
    <reg>Как</reg>
  </choice>
  <choice>
    <orig>вѣтеръ</orig>
    <reg>ветер</reg>
  </choice> черный, поешь: ты мой!
</l>
```

Additions, deletions, substitutions

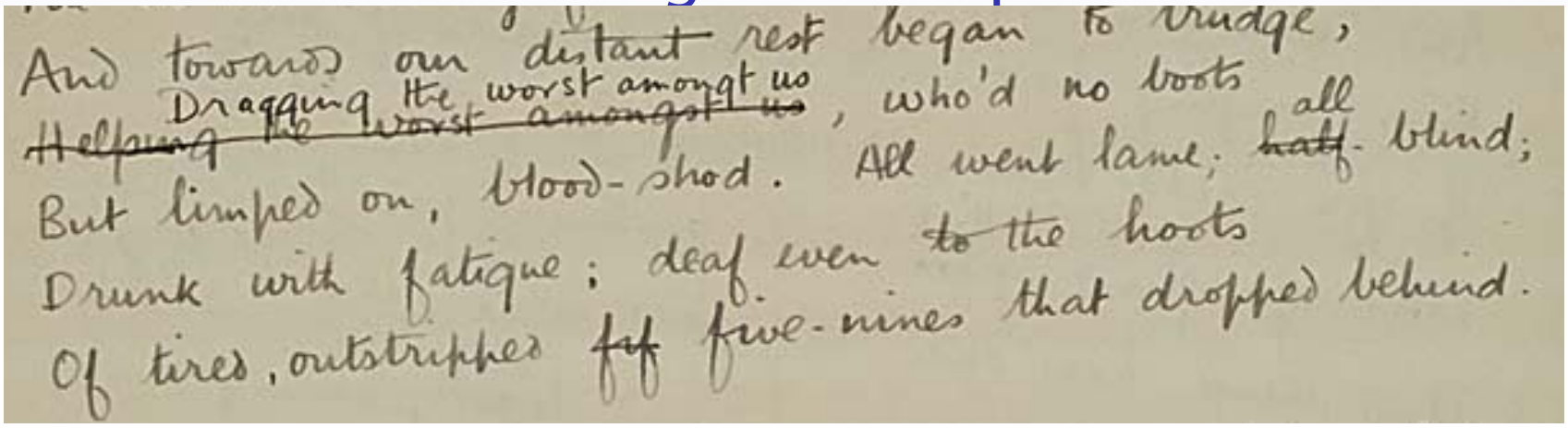
Alterations made to the text, whether by the scribe or in some later hand, can be encoded using <add> (addition) or <del> (deletion).

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




```
<subst>
  <del>half-</del>
  <add>all</add>
</subst> blind
```

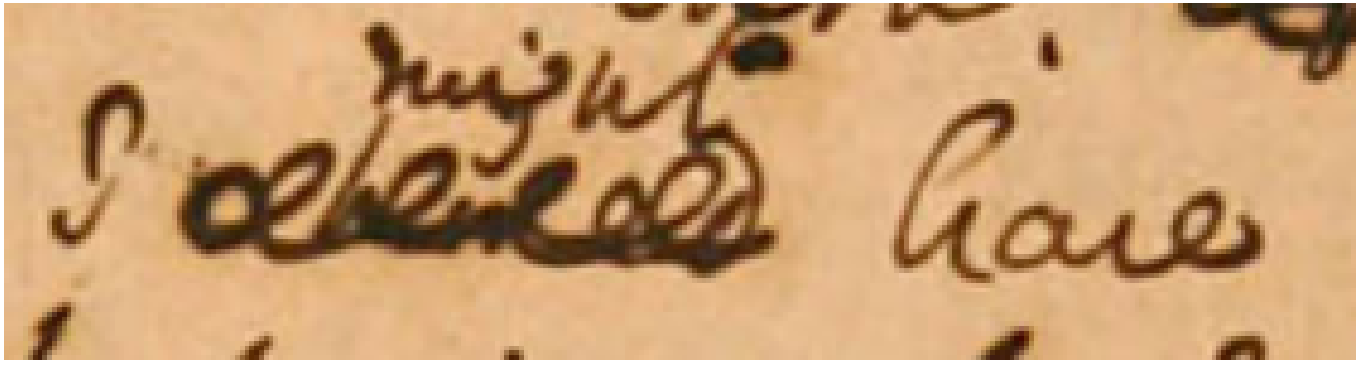
An English example




```
<l>And towards our distant rest began to trudge,</l>
<l>
   <subst>
      <del>Helping the worst amongst us</del>
      <add>Dragging the worst
            amongt us</add>
   </subst>, who'd no boots
</l>
<l>But limped on, blood-shod. All went lame; <subst>
      <del    status="shortEnd">half-</del>
      <add>all</add>
   </subst> blind;</l>
<l>Drunk with fatigue ; deaf even to the hoots</l>
<l>Of tired, outstripped <del>fif</del> five-nines that dropped behind.</l>
```

Semi-legible text

Use <unclear> if the text is partly illegible i.e. it can be read but without perfect confidence.



```
I <subst>
   <add   place="above">might</add>
   <del>
      <unclear   reason="overinking"
         cert="medium"   resp="#LDB"> should</unclear>
   </del>
</subst>have
```

Damage to the carrier

Use the <damage> element to record the existence of physical damage to the document, whether or not the damaged text is readable :

```
<damage   reason="illegible"
  quantity="3"   unit="word"/>Sydney Smith
```

```
<l>The Moving Finger wri<damage   agent="water"   group="1">es; and</damage>
having writ,</l>
<l>Moves
<damage   agent="water"   group="1">
     <supplied>on: nor all
          your</supplied>
  </damage> Piety nor Wit</l>
```

The *@group* attribute is used to associate together parts of the transcription affected by the same area of damage.

Lacunae

Use the <gap> element when something is missing, for example because it is impossible to transcribe, because the carrier is damaged, or because of editorial policy.

In addition to the attributes already mentioned, *@quantity* and *@unit* are available to indicate the size of the gap

```
I am dr Sr yr <gap   reason="illegible"   quantity="3"
  unit="word"/>Sydney Smith
```

```
Their arrangement with respect to Jupiter and to each other was as follows:
<gap   reason="sampling"
  extent="restOfPage">
  <desc>astrological figure</desc>
</gap>
That is, there were two stars on the easterly side and one to the
west; …
```

Text made semi-legible through damage

These elements may be used in combination as necessary. In this example, two phrases can be read despite smoke damage, but three lines in between are completely illegible:

```
<damage agent="smoke">
  <unclear>and the proof of this is</unclear>
  <gap quantity="3" unit="line"
    cause="smokeDamage"/>
  <unclear>margin</unclear>
</damage>
```

Supplied text

Use the <supplied> element if the transcriber has provided a reading not actually visible in the text, perhaps because of scribal error :

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

Attributes can be used to qualify the information further:

- @*reason*why the text has had to be supplied (any word)
- @*source*source (if any) from which the text was taken (a pointer)
- @*resp* who is responsible for supplying this markup (a pointer)
- @*cert*the degree of certainty associated with the markup (high, medium, or low)

Text supplied to fill a lacuna

If the transcriber wishes to supply material to fill a lacuna, it should be marked up using <supplied> rather than <gap>. In this example, 42 lines missing from the bottom of page 5v have been supplied from another source (identified here simply as ed) on the authority of #djb:

```
расѣдша
се. и ѡ
<lb  break="no"/>брѣтени* быти голоубици
<supplied  resp="#djb"  source="ed"
  reason="missing">
  <pb  n="301c"/>
  <lb/>(голоубици) онои въ чре
<lb  break="no"/>вѣ ѥго. и простьръ роу
<lb  break="no"/>коу блаженыи възѧ
<lb  break="no"/>тъ ю живоу. не имоущу
<!-- ... -->
  <lb/>д̄нь и нощь молѧ б̄а о нѥ
<lb  break="no"/>и. и по дъвою лѣтоу вѣ
<lb  break="no"/>сть бысть (ѥмоу. къ-)
</supplied>
<pb  n="6r"/>
<lb/>ѥмоу. гдѣ ѥ<hi  rend="sup">с͡</hi> и како живе
<lb  break="no"/>ть. и оумоливь нѣкого оу
<lb  break="no"/>жикоу своѥго посла тамо.
```

Some difficulties

These methods are perfectly adequate where variation is comparatively simple. They rapidly encounter problems when:

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

How far will the TEI take us ?

In particular, is the TEI scheme adequate for the needs of those transcribing 'modern' manuscripts ?

- surviving medieval or early modern manuscripts generally have a public function, and a more or less conventionalised (if complex) format
- modern manuscripts or authorial drafts however often contain entirely private or idiosyncratic signs, with no clear communicative function

Exercise 1

- Decide on the structure of your encoding
- Decide what you will capture in the transcription and what you will not
- Mark it up!

Look at these files:

- ocsPage.jpg page image
- ocsPage.txt plain text only
- Detailed instructions