

# 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

2-309

Въ тогды дѣла суботный, а  
когда солнце жинило, созидаша  
ѣбо плещи. Поны же и чюдъ  
прѣбывае. Свѣтлый свѣтъ  
поны, и быша тут сѣна зла и  
вѣсѣла с мѣстныи и чюдныи  
сѣна трѣсъ от лѣнныи  
коныи и зѣны от узоръ  
мечей, и казаша, что сѣно  
завершено дѣлашася, и се  
было видно зѣло, ѣбо все  
было залито кровью.



Въ тогды дѣла суботный,  
свѣтъ посредно и  
свѣтлымъ ономъ поны.  
Поны же и чуда  
прѣбывае. Свѣтлый свѣтъ  
поны, и бысть тут сѣна зла  
и вѣсѣла поныи и чюдныи  
сѣна трѣсъ от лѣнныи  
коныи и зѣны от узоръ  
мечей, и казаша, что сѣно  
завершено дѣлашася, и се  
было видно зѣло, ѣбо все  
было залито кровью.

Дѣлашася въ тогды дѣла суботный, а  
когда солнце жинило, созидаша  
ѣбо плещи. Поны же и чюдъ  
прѣбывае. Свѣтлый свѣтъ  
поны, и быша тут сѣна зла и  
вѣсѣла с мѣстныи и чюдныи  
сѣна трѣсъ от лѣнныи  
коныи и зѣны от узоръ  
мечей, и казаша, что сѣно  
завершено дѣлашася, и се  
было видно зѣло, ѣбо все  
было залито кровью.

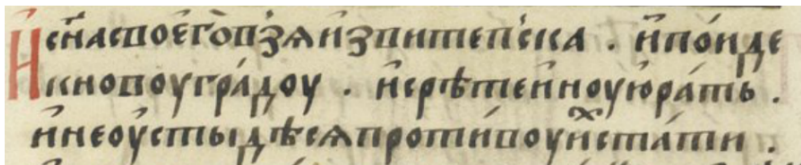
И в синахъ естъ описанно,  
расположенно имя, что воиска  
иже быша в мѣстѣ,  
прѣбывае на немъ. Поны же  
и чуда прѣбывае. Свѣтлый  
свѣтъ поны, и быша тут сѣна  
зла и вѣсѣла с мѣстныи и  
чюдныи сѣна трѣсъ от лѣнныи  
коныи и зѣны от узоръ  
мечей, и казаша, что сѣно  
завершено дѣлашася, и се  
было видно зѣло, ѣбо все  
было залито кровью.

Се же синахъ естъ описанно,  
расположенно имя, что воиска  
иже быша в мѣстѣ,  
прѣбывае на немъ. Поны же  
и чуда прѣбывае. Свѣтлый  
свѣтъ поны, и быша тут сѣна  
зла и вѣсѣла с мѣстныи и  
чюдныи сѣна трѣсъ от лѣнныи  
коныи и зѣны от узоръ  
мечей, и казаша, что сѣно  
завершено дѣлашася, и се  
было видно зѣло, ѣбо все  
было залито кровью.

<http://www.akteon-elib.ru/LITsleVOI-LIeTOPISNYI-SV/D-Russkaya-letopisnaya-istoriya-Kniga-6/#24>



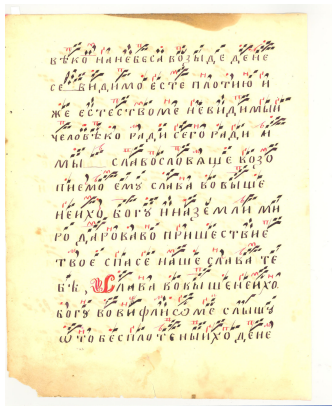
## Transcription is not an automatic process



**Н** сына своего взя из  
Витепска, и поиде к  
Новуградѣ, и сретѣ инѣю  
рать, и не ѹстыдѣся  
противѣ ихъ 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](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*

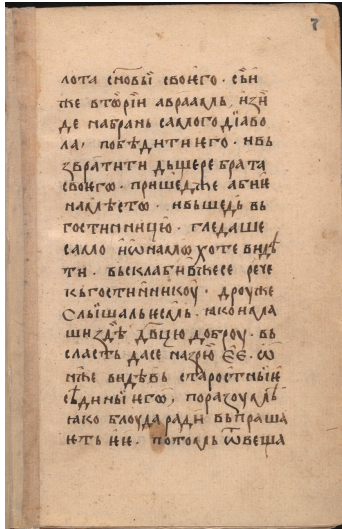
Does your transcription represent a 'text' or a 'document' ?

# How is your transcription organised?

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

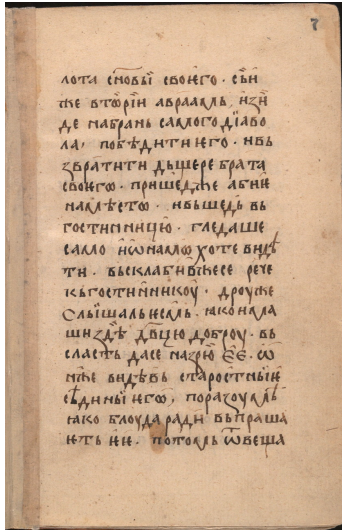


# Layers of transcription



- 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

# minimal encoding



<lb/>лота снѡвы своѣго. съи  
<lb/>же вт<g ref="#oolig">oo</g>рїи авраамъ, изи  
<lb break="n"/>де на брань самого діаво  
<lb break="n"/>ла, побѣдити єго. и въ  
<lb break="n"/>звратити дьщере брата  
<lb/>своєг<g ref="#oolig">oo</g>. пришед же абие  
<lb/>на мѣст<g ref="#oolig">oo</g>. и вышедъ въ  
<lb/>гостинницу. глѣдаше  
<lb/>само и внам<g ref="#oolig">oo</g> хоте видѣ  
<lb break="n"/>ти. въсклабив же се рече  
<lb/>къ гостиннику. друже  
<lb/>слышалъ єсмь. како има  
<lb break="n"/>ши здѣ двѣцу доброу. въ  
<lb/>сласть да се назрю єе. w  
<lb break="n"/>н же видѣвъ старостнык  
<lb/>сѣдины єг<g ref="#oolig">oo</g>, поразоумѣ  
<lb/>како блѡуда ради въпраша  
<lb break="n"/>ють нѣ. потѡмъ ѡвѣща

## 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, columns, 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

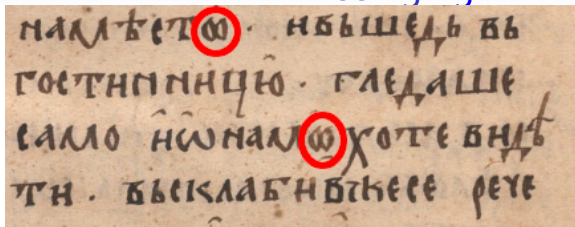
- e.g. **ɑ ɑ ɑ ɑ ɑ ɑ ...** ==> U+0061

- e.g. **Œ** ==> U+0073 **ℒ** ==> U+017F

- 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">oo</g>. и вышедъ въ  
<lb/>гостинницю. глѣдаше  
<lb/>само и ѿнама<g ref="#ooLig">oo</g> хоте видѣ  
<lb break="no"/>ти. въсклабѣнъ бѣ се рече
```

#ooLig points to a description of the glyph, provided in the TEI 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'

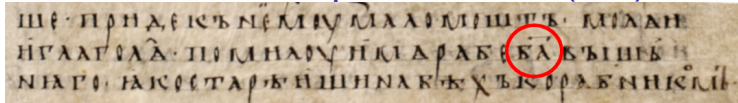
## Two sets of tags

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



## Codex Suprasliensis 61r (119)



We can expand the entire abbreviation in the middle line above with `<abbr>` and `<expan>`:

```
<p>  
  <lb/>ѣ глагола· помилоуѣ ма рабе <choice>  
    <abbr>ба</abbr>  
    <expan>бога</expan>  
</choice> въшѣ-</p>
```

We can use `<abbr>` or `<expan>` alone, without the `<choice>` wrapper, if we want to record just the abbreviation or just the expansion.

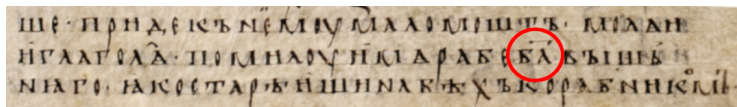
## <choice> is good for more than just abbreviations

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

- <choice> (groups alternative editorial encodings)
- Abbreviation:
  - <abbr> (abbreviated form)
  - <expn> (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>)

## Codex Suprasliensis 61r (119)



Instead of giving the full abbreviated and expanded forms, we can tag the parts with `<ex>` and `<am>`:

<p><lb/>й глагола̑· помилоуи̑ ма рабе б<choice><am>̑</am><ex>ог</ex></choice>а вѣшь-</p>

Here we tag the title as `<am>` ('abbreviation mark') and the expansion added by the editor as `<ex>`.

## 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="brevigraphie">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 finir [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 be combined within a `<choice>` element:

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

## Different regularization strategies (1)

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

Since orthographic regularization (for example for retrieval purposes) is not always predictable :

A transcriber may elect to regularize silently...

<l>Как ветер мокрый, ты бѣешься в ставни,</l>

... or to indicate which words have been regularized:

<l><reg>Как</reg> <reg>ветер</reg> мокрый,  
ты бѣешься <reg>в</reg> ставни,</l>

## Different regularization strategies (2)

... or to indicate both regularized and original forms:

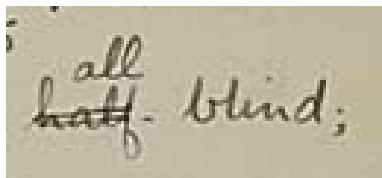
```
<l><choice>
  <orig>Какъ</orig>
  <reg>Как</reg>
</choice>
<choice>
  <orig>вѣтеръ</orig>
  <reg>ветер</reg>
</choice> мокрый, ты бьешься
<choice>
  <orig>въ</orig>
  <reg>в</reg>
</choice> ставни,</l>
```

The same techniques may be used for corrected (<corr>) and erroneous (<sic>) forms

## 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

And towards our distant rest began to trudge,  
~~Helping the worst amongst us~~ Dragging the worst amongst us, who'd no boots all  
But limped on, blood-shod. All went lame; half-blind;  
Drunk with fatigue; deaf even to the hoots  
Of tired, outstripped ~~five~~ five-nines that dropped behind.

<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>

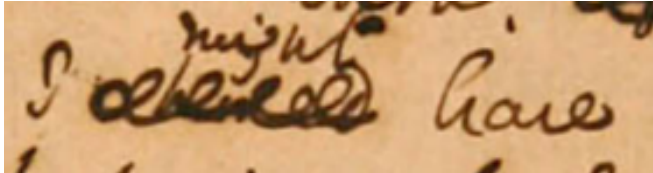
</subst> blind;</l>

<l>Drunk with fatigue ; deaf even to the hoots</l>

<l>Of tired, outstripped <del>five</del> five-nines that dropped behind.</l>

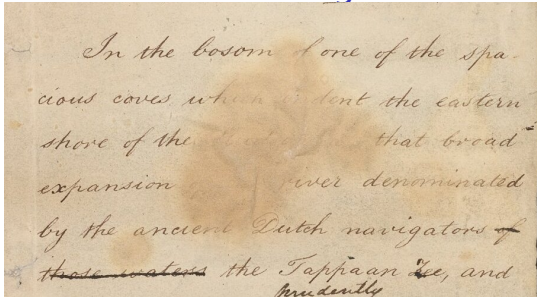
## Semi-legible text

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 :

IN the bosom `<damage group="1">o</damage>`f one of those spa `<lb n="2"/>`cious coves wh`<damage group="1">ich`  
inde`</damage>`nt the eastern `<lb n="3"/>`shore of the  
`<damage group="1">`Hudson, at `</damage>`that broad  
`<lb n="4"/>`expansion `<damage group="1">`of the  
r`</damage>`iver denominated `<lb n="5"/>`by the  
ancie`<damage>`nt`</damage>` Dutch navigators

The @group attribute is used to associate together breaks of the

## Supplied text (1)

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

```
...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)

## Supplied text (2)

Here, “#djb” has decided to fill a lacuna in the source being transcribed using material from edition “ed23” :

```
<lb break="no"/>воу свою, и расѣдша се. и ш  
<lb break="no"/>брѣтени* быти голоубици  
<supplied resp="#djb" source="#ed23" reason="missing">  
  <pb n="301c"/>  
  <lb/>(голоубици) онои въ чре  
  <lb break="no"/>въ юго. и простъръ роу  
  <!-- ... -->  
  <lb/>днь и ноць мола ба о нѣ  
  <lb break="no"/>и. и по дъвою лѣтоу въ  
  <lb break="no"/>сть бысть (юмоу. къ-)  
</supplied>  
<pb n="6r"/>  
<lb/>ѣмоу. гдѣ ю<hi rend="sup">с</hi> и како живе  
<lb break="no"/>тъ. и оумоливъ нѣкого оу  
<lb break="no"/>жигоу своѣго посла тамо.
```

## Supplied text (3)

Alternatively, a `<gap>` element could be used to show that text is missing:

```
<lb break="no"/>брѣтени быти голоубици  
<gap quantity="42" unit="line"/>  
<pb n="6r"/>  
<lb/>ємоу. гдѣ ю<hi rend="sup">с</hi> и како живе  
<lb break="no"/>тъ. и оумоливъ нѣкого оу  
<lb break="no"/>жикоу своѣго посла тамо.
```

## Editorial phrase-level elements

A summary list of some of the more important phrase-level transcription elements might include:

- Core module: `<abbr>`, `<add>`, `<choice>`, `<corr>`, `<del>`, `<expan>`, `<gap>`, `<orig>`, `<reg>`, `<sic>`, `<unclear>`
- 'transcr' module: `<am>`, `<damage>`, `<ex>`, `<metamark>`, `<mod>`, `<redo>`, `<restore>`, `<retrace>`, `<space>`, `<subst>`, `<supplied>`, `<surplus>`, `<transpose>`, `<undo>`

## Some difficulties

These methods are adequate for simple cases but 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

The TEI offers additional features for transcription of modern manuscripts, in which these problems are particularly common.