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 NUMBER NAME OF VERDOR мечей; и казалось, что озерн замерошее двинуюсь, и не GAUDO RECERCIDADA DESO REC. было залито кровью.

BE YOUAA AREA SYNOTHIAR COMPREY ROCYCLINING II

IRANIA II RENTE TV OFMA 344

IS DEAMED HEARDEN IN MOAN, IS

TOYS, OT KONEN ADMARTHE IS

ADDITION IN THE REAL AND THE MEAN, HORDMAN HE COT'L HOR

MAK OT MERINARD DESERTING.

RECISE MODE HOMENIUS

COORDO

HOMIN SEE II YOAN INDERSTINATE CRIMINARY COLORS цин . нетветтеутишлоконплацы . настий нутопоокнийленинаетескоекполин. на в moverkyazannena nokannanyiern . Hmes WIGHTENARMAEHIE . H 3 HOYICTE WMETHAPOCTETE nin . TAKOSKEMOPTOTIOMPTS JUIOLUHTHOVITHEA . инек киндактила Хпонрыловобисекропыто:

TAROMORIA ANACOVEOMINANA . CANYOVHOCYOLA

AFRIEADILA TRIVEAM COUNTY . PETEMHIMIOUH LIKYERATAKH ERKIAHALLOZIOVET TOMULELLE HATTOMOQUETEAHICOMOVICHSHAAIZAHAOV . H REFELMBIGESKIEHCHAOH . HEITHACALHH . HETTH MYHICOVEOGHEAHEATEEA . CIOFEGOALIIKOODBEILOIO Прелилия. налимелтниналецинской. нет KAYOV MERONAMIAKHIOA200V. HHEE'KHANTEK Досоуколнульскогокерега. ниаления ATTELNETETHEATOS MHUSKACITUO . ADIVICAMH TAWANTKMEYA . H . MOVING . HAPOTHINGMOV жененлныутаворий. Нивониетопувано regula . Annurranolanomonu . Anningat MannekaluAnwertroutA . Ratemaskeronen AITAHAA. RTA. F. HATTOYTIAAOVETTIALAFUA HARAMAMILEMITOMINICARAMILIANA . 2 Lik SKEILS OF AAUHET AU FAHRETORINS AAAERCEAN DA Пропетамиполин . Миссислиланилоусон ХНИЛ - СЕЗКЕРЕКЛНИЧКАЦЫ - НАКАТЕЦЕЛНИЕ POICHS AAAEKCANA PAPOVICAMI HEREMOVERTANDEALEMA apoyurkero . nnew KarkIIIICAMAO EMSTERAHI HILLON MAN , MATTANANA CENTRAL EXCLOSION IN DAMAGE CENTURY ACCOUNTS KAAR, MAR AP HIP INCOMEN FOR REALING Алектандал пред всели пенни nce House Hassana y Conven or we serve assess offerer REALING ASSOCIATION OFTIMES, eng ener har speasers a perce are I BE OFFICER OFFICER SALVER

и Глеба, которые за нее пролисни no vóstro 500 mennes, a vyes KIARS 50 DESTRICT WYROF, STREEMS у Нерялова, «Вод – говержия BA ISOMANA, SCANSOWS' SPERIOD AMERICANUES, N. DECEMOR DESIGNATION снало, и склучал Фофин и склучно ACCOUNTY ROUGE & CAMER. COLOR AREA OFTERS EARLING CODA. IN BO IGO HAS KAND L'TOTADA, M CHILLA CONFRICTION INPERA, H BAN BRANCEL STOP, & NOLAH EXCHANGE AND STREET, SO MICH. HADDINGTH WORD'S CHARLY GOODAL H педона ну в Новгород, а нима

http://www.akteon-elib.ru/LITsleVOI-LIeTOPISNYI-SVOD-Russkayaletopisnaya-istoriya-Kniga-6/#24



Transcription is not an automatic process

HALLO егопалияниенска

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

в вко наневеса возвіде дене селенанто есте плотию и же естествоте невидимин ЧЕЛОВЕКО РАЛИ СЕГОРАЛН А Мы славословаще возо ITHEMO EME CANEA BOBLILIE ห้ยมรับ ธังการ ที่หลังยุภาพี MH Po ภูลิคอีถิสิธิอ กคิมเยียงารที่หย TBOE CHACE HAME CAABA TE Б.К., Слава вокы шененхо GOTS BOBH WAH COME CALINS Отобесило тенинхолене

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

TEI

 all transcription is imaginative

MS5045 Russian Musical Manuscript, from

^{5/34} http://faculty.goucher.edu/eng241/high_church_slavonic_leaf_images.ium

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

5 r ferry for Enrgane de -Proante Popt proved an mare a Vug felin a servir de zemede conte coarme puer po? Envenate foubs formen . Guz le prise De grinzo Cinzos town Le mair Targent Lerop fuff cofines ca faburation DEre to four a Die Deniene Dro Brit graine frous quarts it fin qui bal buzo seriore ho graine Daroyent (2007, a.

<pb n="13" xml:id="B452346101_Ms629_0015"/> <fw place="top-right" type="pageNum">5</fw> <lb/>Samp; demy trebuchans de <lb/>soixante sept pieces au marc <lb/>a ung felin & demy de remede <lb/>Cours chacune piece <expan>pour</expan> <lb/><expan>cinguante</expan> soubz <expan>tournois</expan><pc>. <lb/><lb/>Sur le pris de quinze livres <lb/><expan>tournois</expan> le marc dargent le rov <lb/>fust <expan>continuee</expan> la <expan>fabricacion</expan> <lb/>des testons a dix deniers <lb/>dix huict grains trovs guartz <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. **Q Q Q A A A A ...** ==> U+0061
• e.g. **S** ==> U+0073
$$\int$$
 ==> 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>

נהסורים . חדרשור אבריני אברות אמאדנידים . אבנשנאני אב רסנידארוארווים . ראנגמשני נמאחם אישאמאש צסידינ באגל דא . בנוגאמדי ביאנינו פיינ

Bdinski, fol 7r, detail

There is no Unicode character for the oo glyph here: we tag it as a $<\!\!g\!>$

#ooGlyph 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



A French example

Conze ce armin form Envquate

We might just note that we have expanded the abbreviations:

<lb>Cours chacune piece <expan>pour</expan><lb/><lb/><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.

```
<abbr>po&#x1D63;</abbr> .... <abbr>cinqu&#x0101;te</abbr>
```



... or we might look a bit closer

.We can tag the abbreviation markers and the expansion directly

po<am>ᵣ</am> ... or po<ex>u</ex>r

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

<abbr>po<am>ᵣ</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 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, 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 бъешься въ ставни, Какъ вѣтеръ черный, поешь: ть: мой! Я досвый хаосъ, в другъ твой давній.

<l>

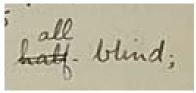
```
<choice>
    <orig>Kaкъ</orig>
    <reg>Kak</reg>
  </choice>
  <choice>
    <orig>вѣтеръ</orig>
    <reg>ветер</reg>
  </choice> мокрый, ты бъешься <choice>
    <orig>въ</orig>
    <reg>B</reg>
  </choice> ставни.
</1>
<1>
  <choice>
    <orig>Kaкъ</orig>
    <reg>Kak</reg>
  </choice>
  <choice>
    <orig>вѣтеръ</orig>
    <reg>ветер</reg>
  </choice> черный, поешь: ты мой!
</1>
```



Additions, deletions, substitutions

Alterations made to the text, whether by the scribe or in some later hand, can be encoded using <add> (addition) or (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> half- <add>all</add> </subst> blind



An English example And towards our distant rest began to trudge, Dragging the worst amongst us, who'd no boots all Helping the more shod. All went lame, hatf. blind; But limped on, blood-shod. All went lame, hatf. blind; Drunk with fatigue; deaf even to the hoots Of tires, outstrippes for five-nines that dropped behind.

<l>And towards our distant rest began to trudge,</l>

<subst>

```
<del>Helping the worst amongst us</del>
<add>Dragging the worst
amongt us</add>
</subst>, who'd no boots
</l>
but limped on, blood-shod. All went lame; <subst>
<del status="shortEnd">shaf-</del>
<add>all</add>
</subst> blind;</l>
Drunk with fatigue ; deaf even to the hoots</l>
</>>
```



Semi-legible text

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

```
I <subst>

<ad place="above">might</add>

<del>

<unclear reason="overinking"

cert="medium" resp="#LDB"> should</unclear>

</del>

</subst>lave
```



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
    Moves
    damage agent="water" group="1">
    supplied>on: nor all
your</supplied>
    </damage> Piety nor Wit</l>
    </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 $\langle supplied \rangle$ rather than $\langle gap \rangle$.

break="no"/>брѣтени быти голоубици <supplied resp="#djb" source="ed" reason="missing"> <pb n="301c"/> <lb/>(голоубици) онои въ чре < --> <lb>break="no"/>сть бысть (кемоу. къ-) </supplied> <pb n="6r"/> <lb/>смоу. гдѣ ю<hi rend="sup">c</hi> и како живе <lb break="no"/>ть. и оумоливь нѣкого оу <lb break="no"/>жикоу своюго посла тамо.

Alternatively, the transcription might simply indicate that the text is missing and not supply it

<lb>break="no"/>брѣтени быти голоубици <gap quantity="42" unit="line"/> <pb n="6r"/> lb/>ємоу. гдѣ ю<hi rend="sup">c</hi> и како живе <lb break="no"/>ть. и оумоливь нѣкого оу <lb break="no"/>жикоу своюго 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

