


<oo> → <bs> *Bdinski sbornik*

Author: David J. Birnbaum (djbpitt@gmail.com)

Maintained by: David J. Birnbaum (djbpitt@gmail.com) 

Last modified: 2014-06-03T07:20:12+0000

Bdinski sbornik encoding guidelines

Developed by David J. Birnbaum, Michel de Dobbeleer, Alexandre Popowycz, and Lara Sels

Contents

- [Introduction](#)
- [Schemas](#) (Relax NG and Schematron)
- [Structural markup](#)
 - Features of the manuscript: [lines](#), [folios](#), [superscription](#), [word division](#), [color](#), [abbreviation](#), [lacunae](#), [corrections](#), [deletions and insertions](#), [damage](#), [generic problems](#)
 - Features of the text: [beginning and end](#)
 - Features of the digital edition: [metadata](#) (root, editor, email)
 - Features of the 1973 print edition: [paragraphs](#), [pages](#), [asterisks and footnotes](#)
- [Transcription and character coding](#)
 - Letters: [capitalization](#), [ligatures](#), [e](#), [o](#), [u](#), [jery](#) and [jer letters](#), [t](#)
 - Non-letters: [titlo and pokrytie](#), [other diacritics](#), [numbers](#), [punctuation](#)

Introduction

This document is a guide for encoding the *Bdinski sbornik* in XML, which is subsequently transformed to HTML for publication at <http://bdinski.obdurodon.org>. The starting point of the digital edition is the 1973 typeset edition, which we digitized through [optical character recognition](#). Each textual unit (text) is encoded as a separate XML file; for example, the raw XML-encoded version of the first (Abraham) text is available at <http://bdinski.obdurodon.org/abraham.xml>. If you just click on this link, some browsers will download the file, some will display it with markup, and some will display just the textual content, without markup. If what you see is not what you want (browsers typically reformat the white space in XML files for rendering, which misrepresents the contents), you can download the file instead of just clicking on the link, whereupon you can open it in the application of your choice. Project development is performed with the help of the [<oXygen/>](#) XML editor and integrated development environment.

Editing with a schema

The structure of a file for the project (which elements are permitted to occur where) is described by a Relax NG schema ([bdinski.rnc](#)) and a Schematron schema ([bdinski.sch](#)), which can be linked to the file itself in [<oXygen/>](#). We use both schemas because they validate different aspects of the structure. There are two reasons to link the schemas to the file:

1. If a schema is linked to the file, [<oXygen/>](#) will perform real-time validation, providing feedback during the editing process about whether the markup follows the rules. This helps

the encoder avoid introducing erroneous markup.

2. If a schema is linked to the file, <oXygen/> will provide prompting and command completion. For example, if the encoder types an angle bracket, <oXygen/> will use the schema to determine which elements are legal at the current cursor position, and will drop down a list of items on which the encoder can click to enter the tags.

To attach a schema to a file, click on “Document” in the menu bar above the <oXygen/> editing window, then on “Schema,” and then on “Associate schema.” To the right of the text-entry box labeled “URL” is a small icon shaped like a folder. Click on that and navigate to the main Bdinski sbornik Dropbox folder, which is called “bdinski-sbornik.” Inside that folder, select the file called “bdinski.rnc,” ensure that “Use relative paths” is checked, and click “OK.” (You don’t have to specify the schema type; <oXygen/> will figure that out on its own.) Then do the same with “bdinski.sch”. For each schema you add <oXygen/> will insert a line near the top of the file you are editing, and with both schema links in place the top of your file will look something like:

```
<?xml-model href="../../../bdinski.rnc" type="application/relax-ng-compact-syntax" />
<?xml-model href="../../../bdinski.sch" type="application/xml" schematypens="http://www.w3.org/2001/XMLSchema" />
```

You have to do this only once for each file you edit. Once the lines have been inserted by <oXygen/> into the top of your document, <oXygen/> will know to use the two schemas to validate the document and to provide command completion.

Markup

Features of the manuscript

The *Bdinski sbornik* markup is based primarily on the physical layout of the manuscript as a series of folios, each of which contains a series of lines. We always encode entire folios, which means that if a text begins in the middle of one folio and ends in the middle of a different folio, we encode both folios in their entirety, including the lines that occur before the beginning of the text we are editing and those that occur after the end of that text.

Lines

Every line in the manuscript is encoded as a <line> element, with a <line> start tag and a </line> end tag, e.g.:

```
<line>ДВЕРЦА МАЛА ПОСРѢДѢ ЕЮ.</line>
```

If a word is divided across a line break, the editors should add a hyphen at the end of the first of the two lines, corresponding to modern hyphenation conventions, e.g.:

```
<line>ЖЕ ЕИ СЕДМИЮ ЛѢТЬ. W-</line>
<line>Н ЖЕ ПОВЕЛѢ ЕИ БЫТИ ВЪ</line>
```

Folios

The beginning of each folio is marked as an empty (self-closing) <folio> element, which has an

@n attribute that consists of a sequence of digits followed by the letter *r* for “recto” or *v* for “verso.” The following example shows an abbreviated version of how the encoding of folio 22r and folio 22v might look:

```
<folio n="22r"/>
<line>first line of 22r</line>
<line>second line of 22r</line>
<!-- more lines -->
<line>last line of 22r</line>
<folio n="22v"/>
<line>first line of 22v</line>
<line>second line of 22v</line>
<!-- more lines -->
<line>last line of 22v</line>
```

Superscription

Superscription is encoded with the **<sup>** element, e.g.:

```
<line>ВЪНЪШНЪИ ХИЗИНЪ. СМ<sup>Ъ</sup></line>
```

Omega with superscript “t” is rendered not as the single Unicode U+047f (Ѡ), but as omega followed by “t” wrapped in **<sup>** tags, e.g.:

```
<line>СЛЪ. И Ѡ<sup>Т</sup>ВРЪЗШИ ДВѢРЦЕ</line>
```

Word division

Word division is entirely editorial, which is to say that the editors insert spaces between all words, regardless of spacing in the manuscript. The particles *ѡ* (reflexive) and *жѡ* are independent words, and are preceded by spaces. This is true even when they are superscripted, so *ѡн ѡ* (with a space before the superscript *ж*) rather than *ѡнѡ* (with the superscript *ж* immediately adjacent to the preceding *н*).

Colored text

Text in red should be tagged as **<red>**. In cases where an entire line may be in red, the **<red >** tags should go inside the **<line>** tags, e.g., 1r2 should be encoded as:

```
<line><red>ЖЕНАГО АВРАМИА. КАКО П-</red></line>
```

Note that **<red>** tags, like **<line>** tags, may capture only the beginning or end of a word.

Abbreviation

The 1973 edition expands abbreviations, wrapping the inserted letters in parentheses. We remove those, restoring a tilde or superscript letter plus pokrytie if there was one in the manuscript.

Lacunae in the manuscript

Some of the text that would have occurred on folios now missing from the manuscript has been restored in the 1973 edition on the basis of other manuscripts. For our purposes, we keep that text, without dividing it into lines, but surround it in **<lacuna>** tags.

Corrections

Errors corrected in the manuscript by a scribe (original or subsequent) are encoded by creating a **<subst>** (= substitution) element. **<subst>** must have exactly two children, one instance of **** (which contains the original reading that was subsequently corrected by the scribe) followed by one instance of **<add>** (which contains the text inserted by the scribe as a correction).

If the original reading is illegible because of the correction, it may be rendered as a **<gap/>** element (with an optional **@extent** attribute to indicate the number of characters, if the editor can discern that with reasonable confidence) inside the **** element. For example, the replacement of one illegible character with two at the end of 39v16 should be encoded as:

```
ѠБРАЗ<subst><del><gap extent="1" reason="overwritten"/></del><add>wm</ad
```

If the text is legible but unclear, it may be wrapped instead in **<unclear>** tags inside the **** element.

In case of corrections that do not involve the complete deletion of an initial value, the **** element should carry the attribute **@status** with the value "partial". For example, in 57r3 the scribe began to write л and then corrected himself to и, and we write:

```
ПОМ<subst><del status="partial">л</del><add>и</add></subst>ШЛІАІѦ
```

This markup is borrowed from the TEI P5 guidelines. The notion of partial deletion reflects the TEI interpretation that **** does not have to mean complete deletion (in the case of the example above, the deletion is conceptual, rather than graphic). Note that the TEI **<corr>** element should not be used to encode corrections that can be read in the manuscript, that is, that were created by a scribe. **<corr>** is to be used only for corrections inserted by the modern editors, and at this stage in our development the editors of this project are not encoding any new corrections.

Deletions and insertions

Text that has been erased and not replaced but that is still legible should be tagged as ****. If the deleted text is not legible but it is clear that text was deleted, the **** element should still be used, but it should contain only an empty **<gap reason="erased"/>** element, with an optional **@extent** attribute to indicate the number of erased letters, if the editor can determine that with reasonable confidence. For example, the five-character erasure at 40v8 should be transcribed as:

```
<line>ВЪ ДООМЬ<del><gap reason="erased" extent="5"/></del>НИСИФОРОВЪ.</line>
```

Do not use **<gap/>** by itself for this purpose; wrapping it in **** is what makes explicit that the

gap results from scribal deletion, and not from **damage** to the manuscript or for other reasons. If the text is partially legible, it is possible to combine raw text and **<gap/>** elements inside a **** element. **<gap/>** optionally may contain an **@extent** attribute that records the estimated number of characters deleted, so that, for example

```
<del>и<gap reason="illegible" extent="2"/></del>
```

records a three-letter erasure where the first letter can be read with reasonable confidence as и and the next two are illegible.

Text inserted into the manuscript by a later scribe should be tagged as **<add>**. If the editor is confident that the insertion is in a later hand, optional **@hand="other"** attribute markup may be included (insertions by the original scribe should omit the **@hand** attribute entirely, since the original scribal hand is assumed to be the default). For example, if the editor is confident that the superscript Δ at 46r12 was added in a later hand, that can be encoded as:

```
соу<add hand="other"><sup>Δ</sup></add>тъ
```

Combinations of deletions and insertions that should be regarded as connected, that is, that should be considered a correction, should be encoded using **<subst>**, as described above, under [Corrections](#).

Physical damage to the manuscript

Text that is legible but damaged should be tagged as **<damage>**, e.g., at 39v17:

```
пΔБ<damage>бΔб</damage>.
```

If the damaged text can be read, but not with confidence, it (or any unclear portions) can be wrapped in **<unclear>** tags inside the **<damage>** element. If the text cannot be read at all, it should be tagged as an empty **<gap/>** element inside **<damage>**, where **<gap/>** has an optional **@extent** attribute that indicates the approximate number of illegible characters (where the editor is able to discern).

Generic problems

Generic problems should be tagged as **<problem>**. These will be resolved and reclassified later, after discussion, and this interim markup will help find them at that time.

Features of the text

The beginning and end of a text

The beginning of the text being edited is marked by inserting an empty **<start/>** tag before the line on which the text begins, and the end of that text is marked by inserting an empty **<end/>** tag after the last line. For example, if the text being edited begins on the third line of folio 22r, the markup would look as follows:

```

<folio n="22r">
<line>first line of 22r</line>
<line>second line of 22r</line>
<start/>
<line>third line of 22r</line>
<!-- more lines -->

```

There should be exactly one `<start/>` and one `<end/>` tag in each file, surrounding the text being edited at the moment. Do not mark the end of the preceding text or the beginning of the following one; that is, in the example above, do not include an `<end/>` before the `<start/>`

Features of the digital edition

The root element and information about the editor

The entire edited section is wrapped in a single `<root>` element. The first subelement inside the `<root>` must be a `<metadata>` element, which contains the `<name>` and `<email>` of the primary person who edited the section. (This assumes that every section will have exactly one editor to be credited officially on the site.) The following example shows the beginning of `thais.xml`, and demonstrates the `<root>` start tag at the beginning of the file, the `<metadata>` element with its `<name>` and `<email>` children, the `<folio>` tag for folio 106v, which is where this text begins, the `<line>` tags for the lines on that folio, and the `<start/>` tag before the first line of the text of the Vita of Thais. This example also includes the `<sup>` element for superscript characters and an `<editionPageNo>` element, about which see below:

```

<root>
  <metadata>
    <name>Alexandre Popowycz</name>
    <email>alexandre.popowycz@ugent.be</email>
  </metadata>
  <folio n="106v"/>
  <line>рцами. и исписахъ ни все стра-</line>
  <line>сти юже имѣ къ вѣсоу борбѣ.</line>
  <line>и къ вбразномуу и змѣю, и</line>
  <line>все мѣлѣвы юе. и поустихъ</line>
  <line>къ вѣ<sup>м</sup> хр<sup>с</sup> тѣмъ<sup>м</sup> съ всею исти-</li>
  <line>ною. сконча же се. стаа моу-</line>
  <line>ченица марина, м<sup>с</sup>ца ѿула</line>
  <line>въ, зѣ. и твореще ни память,</line>
  <line>спсение оулоучимъ. въ име</line>
  <line>га нашего іс хъ. емѣже сла<sup>в</sup></line>
  <line>и дръжава съ безначелнимъ</line>
  <line>его ѿцемъ. и съ прѣстѣмъ</line>
  <line>и бл҃гимъ и животворещимъ</line>
  <line>дхомъ твоимъ, ѿнѣ и прѣсно:—</line>
  <start/>
  <line><editionPageNo n="130"/>ЖИТИЕ И ЖИЗНЬ ПРѣ-</line>
  <line>по<sup>д</sup>вниѣ дасиѣ,</line>
  <line>Братина моя пр<sup>с</sup>наа, хоцѣ ва<sup>м</sup></line>
  <!-- text continues -->

```

```
</root>
```

Features of the 1973 print edition

Paragraph numbering

Some of the texts in the 1973 edition number paragraphs or larger sections. This information is retained as empty (self-closing) `<editionParagraphNo>` elements with an `@n` attribute to indicate the numerical value, with a space before and after, e.g.:

```
<line>Λβι. <editionParagraphNo n="28"/> ρε<sup>γ</sup> же въ себѣ азъ оуже</line>
```

Page numbers

Page numbers from the typeset edition are not present in the OCR output, and must be inserted manually into the XML by the editors. The markup for this purpose is an empty (self-closing) `<editionPageNo>` element with an `@n` attribute, the value of which corresponds to the beginning of a page in the 1973 edition, e.g.:

```
<line>неже <editionPageNo n="44"/>бо бѣ ѡ<sup>т</sup>цѣ єѣ, имѣ-</line>
```

Asterisks and footnotes

Asterisks in the 1973 edition point to a wide variety of editorial footnotes. When correcting the OCR, the asterisks should be retained during editing, but in places where they indicate that the editors have replaced the actual manuscript text with their own emendation, the actual manuscript text (in the footnote in the typeset edition) should be restored to the transcription.

Transcription and character coding

Letters

Capitalization

The 1973 edition capitalizes proper nouns, sentence-initial letters, etc. We correct those according to the manuscript, which means that we use capital letters only for letters that are large in the manuscript, typically initials and the title at the beginning.

Ligatures

Ligatures are wrapped in `<lig>` tags, e.g.:

```
ρ<lig>αγ</lig><sup>Δ</sup>ε (13v2)
```

Broad and narrow "e" (ε ~ €)

The manuscript distinguishes broad and narrow “e”, but these are not distinguished in the 1973 edition. We correct this programmatically after OCR to bring the distribution into agreement with the general scribal orthographic norm, writing narrow “e” (e) after consonant letters and broad “e” (ε) elsewhere, that is, after vowel letters and in initial position. Because the scribe may occasionally violate his own norm, editors of the new digital edition need to proofread especially carefully for such deviations and correct the transcription, so that it comes to represent the actual spellings in the manuscript.

“o” letters (o ~ w ~ o ~ o ~ oo)

The manuscript includes omicron (o), omega (w), ocular “o” (o, e.g., 6v16), binocular “o” (o, e.g., 5r12) and broad “o” (O, e.g., 8v8). The double omicron (oo) is a distinctive feature of the orthography of this manuscript, and is transcribed as a sequence of two regular omicron letters. The two omicrons typically are touching, and therefore technically a ligature, but we’ll add the ligature markup automatically after editing, so the editors do not have to type the tags manually in these cases.

“u” letters

The sound [u] may be spelled as omicron plus “u” (this should be encoded as two characters, regular “o” followed by regular “u”, e.g., beginning of оумершѣ 1v4), ѣ (e.g., end of оумершѣ 1v4), and with superscript y over omicron, i.e., o^y (e.g., o^yловити 3r3). Note that this means that we have a **<sup>** element within a **<sup>** element in the case of вѣроуом^{o^y} 6v15:

```
вѣроуом<sup>o<sup>y</sup></sup></sup>
```

Jery and jer letters

Jery in the manuscript is regularly written with two marks over the second component, which sometimes looks like two dots and sometimes like a kendema (double grave accent). In all cases we render the jery like regular jery, with front jer onset and without any superscript diacritic (cf. below concerning [diacritics](#)): ѣ.

All jer letters are written as front jer (ѣ) unless they are unambiguously back jers (ѣ), which normally occurs only at the ends of lines.

“t” letters

There are three basic shapes of the “t” letter: three-legged (e.g., тѣо 4r5), regular (but with strong serifs on the ends of the crossbar, e.g., тѣо 4r4), and a tall “t” (e.g., тѣо 4r12). These are all transcribed, identically, as regular т.

Non-letters

Titlo and pokrytie

Titlo is transcribed over the last continuous letter of the word (counting from the beginning) before the first omission, without regard to where it appears graphically in the manuscript. In this way, the titlo records the fact that there is a titlo in the manuscript, but does not attempt to record

its exact placement over a particular base character. This accommodates the fact that titlo may be placed not only over individual letters, but also between letters, and that it may span multiple letters, all of which are features that are impractical to represent in a character-based transcription. The imperative to transcribe with exact placement is relaxed because we provide also photographs, so that users who require this level of paleographic detail have access to it by way of the photographs. Thus:

- $\Gamma\tilde{\Lambda}\Sigma$ (13r9) for $\Gamma\Lambda(\Delta\Gamma\Lambda)\Delta\Sigma$ because, in keeping with usual abbreviation convention, the Δ that is present in the manuscript is most likely the second one, so the first omitted letter is the first Δ , which means that the last consecutive non-omitted letter from the beginning of the word is the first Λ of the word.
- $\tilde{\Pi}\tilde{\Sigma}\tilde{\Gamma}\tilde{\chi}$ for $\Pi(\epsilon)\tilde{\Sigma}(\epsilon)\tilde{\Gamma}\tilde{\chi}$, even in situations where the manuscript may read $\tilde{\Pi}\tilde{\Sigma}\tilde{\Gamma}\tilde{\chi}$ (with titlo over the $\tilde{\Sigma}$), because the last consecutive non-omitted letter from the beginning of the word is Π . In this respect the transcription reflects that the word is abbreviated and contains a titlo, but it does not attempt to reproduce the placement of the titlo. In this particular case, the titlo that occurs over the $\tilde{\Sigma}$ in the manuscript extends also over the following $\tilde{\Gamma}$, but we nonetheless write it only over the initial Π .
- We write only one titlo per word, even if the manuscript contains more than one, so $\tilde{\Pi}\tilde{\Sigma}\tilde{\epsilon}$ for $\Pi(\epsilon)\tilde{\Sigma}(\epsilon)\tilde{\epsilon}$ even where the manuscript might read $\tilde{\Pi}\tilde{\Sigma}\tilde{\epsilon}$.
- We write Jesus Christ as two words, with a titlo over each part, e.g., $\tilde{\iota}\tilde{\epsilon}\tilde{\chi}$ (75r16).

Exception: if the last consecutive non-omitted letter from the beginning of the word is superscripted, we place the titlo over the following letter. Thus $\tilde{\omega}\tilde{\tau}\tilde{\psi}$ for $\omega\tau(\psi)\psi$.

We write pokrytie over superscript letters where it appears in the manuscript. This manuscript does not strongly distinguish titlo and pokrytie by ductus (that is, the squiggles look similar); we use titlo (\sim) exclusively over in-line base characters and pokrytie ($\hat{\sim}$) exclusively over superscript characters.

Other diacritics

Accent marks and other diacritics (other than [titlo and pokrytie](#)) are omitted.

Numbers

Numbers are rendered as they appear in the manuscript, that is, as Cyrillic letters, often with titlo and often preceded or followed by a comma or mid-dot. Numbers in this manuscript are always preceded by punctuation, even where it is not required syntactically, and the punctuation is written adjacent to preceding letter, followed by a space, followed by the number, e.g., $\Delta\tilde{\Pi}\tilde{\iota}$ (with the dot next to the preceding Π and followed by a space), rather than $\Delta\tilde{\Pi}\tilde{\iota}$ (with the space before the dot, which is then adjacent to the following ι). The decisive example is $\tilde{\epsilon}\tilde{\omega}\tilde{\gamma}$. | $\tilde{\epsilon}$. $\Lambda\tilde{\Sigma}\tilde{\tau}$, 2v11–12, where a line break clearly associates the dot with the preceding word, and not with the following number.

Punctuation

The 1973 edition uses punctuation supplied by the editors. We remove that, encoding punctuation as it occurs in the actual manuscript. See the section about [numbers](#), above, concerning punctuation adjacent to numbers.