

Цифровые гуманитарные технологии: ресурсы, инструменты, кейсы. Лекция 2. Оцифровка текстов: стандарты, возможности, ограничения

Борис Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

8 февраля 2018

Содержание

- 1 Оцифровка текстов: обзор
- 2 Оцифровка текстов: концептуальная сторона
- 3 Разметка

Содержание

1 Оцифровка текстов: обзор

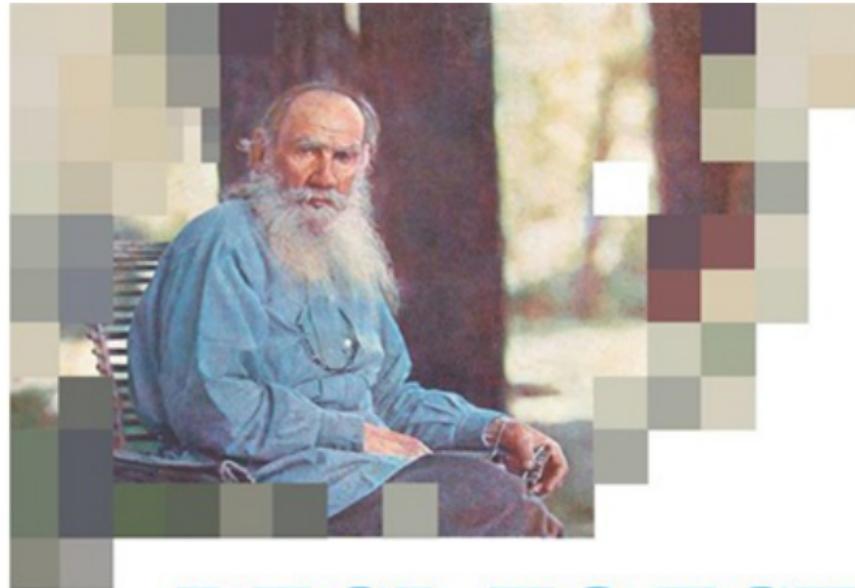
2 Оцифровка текстов: концептуальная сторона

3 Разметка

Оцифровка — это кодирование



Отсканировать недостаточно



Зайдите на сайт
читаемтолстого.рф
и участуйте в проекте!

ВЕСЬ ТОЛСТОЙ В ОДИН КЛИК

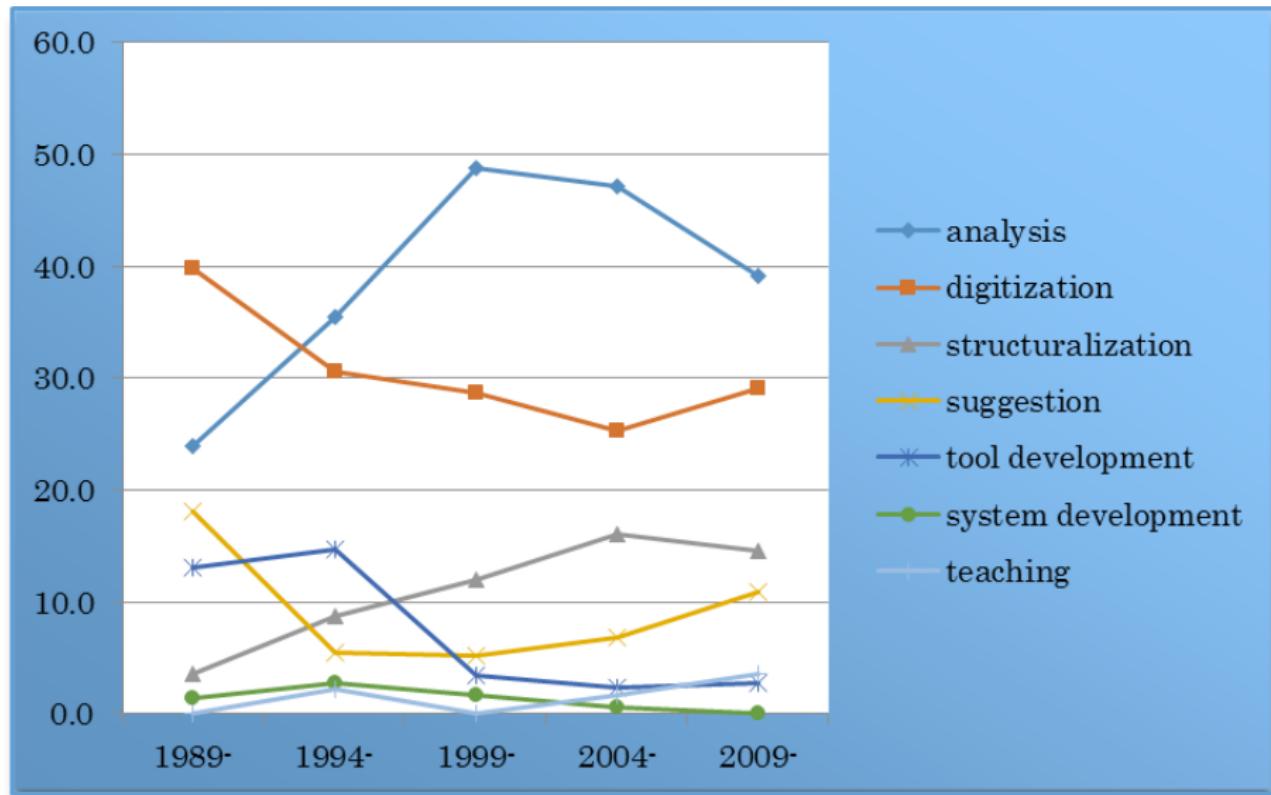
Фильм «Стартап», 2014

Художественное преображение истории «Яндекса».

0:32:46: «Сейчас в систему загружено полное собрание сочинений Льва Николаевича Толстого»

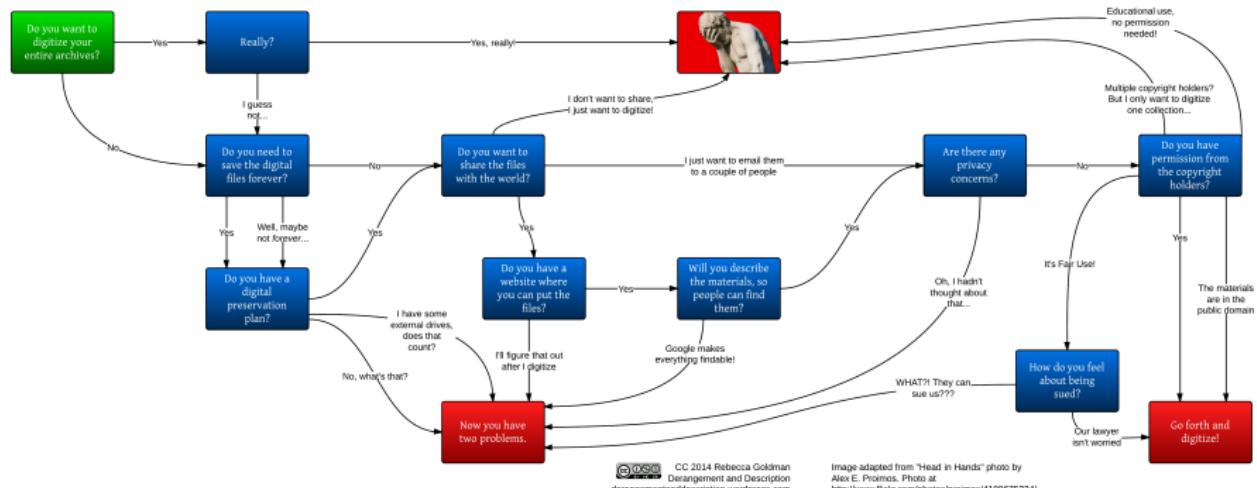
На самом деле в своей ранней истории в «Яндексе» работали над поиском по текстам А. С. Грибоедова.

Оцифровка в трендах DH



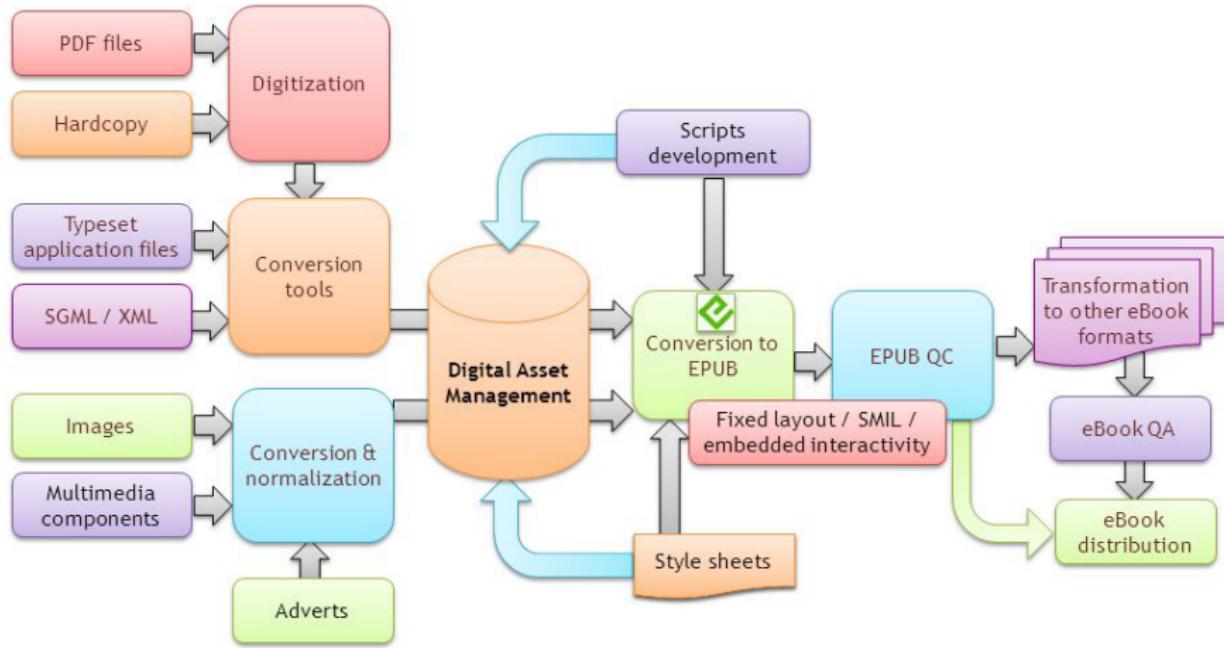
Оцифовка – это проблема

Will digitization solve my problem?



MPS eBook Workflow

MPS



Оцифровка

- Неизбежна: переход человечества на цифровые носители
- Сохранение культурных артефактов
- Удешевление копирования, предоставления доступа к информации
- Электронные публикации могут включать в себя больше документов
- Появляются новые социальные практики (например, shared knowledge)

Роман с камнем: эпиграфика



Decretum de Minervae Victoriae Sacerdoce Temploque (II)

Decretum de Minervae Victoriae Sacerdoce Temploque (III)

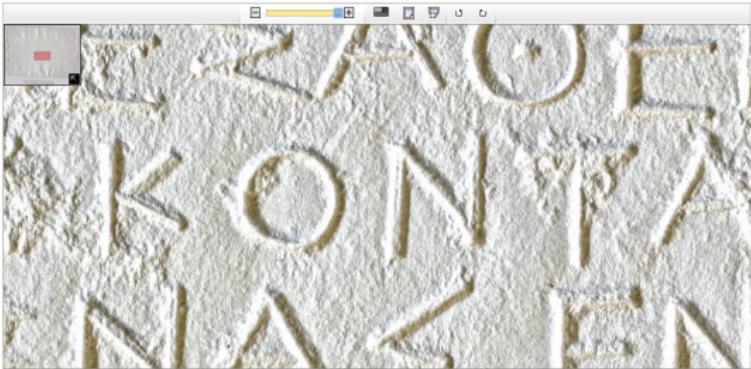
[View Description](#)

Save page

1 of 1

[Download](#)

[Print](#)



Текст vs. документ

With this affair along, for all our thanks,
Now follows, that ~~you know~~ young ~~Ferdin.~~, ~~not~~ ~~but~~ ~~trust~~
~~Holding a weak upper hand over us,~~
~~Or thinking by our late dear Brothers death,~~
~~Our State to be disjointed, and out of Frame,~~
~~Colleagued with the dream of his Advantage;~~
~~He hath not fail'd to pester us with Message,~~
~~Impouting the surrend're of these lands~~
~~Left by his Father, with all Bonds of Law~~
~~To our most valiant Brother, So much for him.~~

Enter Valimund and Cornelia.

Now for our selfs, and for this time of meeting
Thus much the business is. We have here
To ~~Norway~~, Uncle of young ~~Ferdinadus~~,
~~Who importuned and bedid, scarcely heare~~
~~Of this his Nephews purpose, to supprese~~
~~His further care herein. In that he Lyes,~~
~~The Lifs, and full provissions are all made~~
~~Out of his subiects: and we here desirous~~
~~You good ~~Cornelius~~, and you ~~Katharina~~,~~
~~For bearing of this growing and old Norway.~~
Giving to you no further personal power, ~~and~~ ~~to~~ ~~you~~
To be ~~with~~ ~~the~~ King, were then the scope
Of these dilated Articles allow:

Exeunt.

Farewell, and let me what command you may.
Kolt. In that, and all things, will I do my duty.
King. We doubt not but you will.

Enter Valimund and Cornelia.

Дуализм
текста как
документа
и текста
как про-
изведения:
цифровое
издание
гораздо
лучше
представ-
ляет и то,
и другое.

Электронные возможности

The Digital Vercelli Book beta

The screenshot shows a digital edition of the Vercelli Book. On the left is a photograph of an open medieval manuscript with two pages visible. The text is written in a Gothic script. A large initial letter 'H' is prominent on the left page. On the right side, the text is transcribed into modern English. The transcription is as follows:

7 **D**wæt ic swefna cýrt secgan wylle hæt mege matte
8 to midre nihte syðjan reord berend reste wunedon.
9 þuhte me þær ic ge jawe yllicre treow onlyft
10 lædan leohste be wunden beama beorhtoþ eall þær
11 beacen wær be gotten mid golde gimmer stodon fegere
12 æt foldan sceatum . fwylce þær fife wæron uppe
13 on þam eaxle ge spanne be heoldon þær engel dryht
14 nef ealle fægere þurh forð ge sceafte newæt ðærhuru
15 fracodeþ gealga . achine þær beheoldon halige gaſta
16 men ofer moldan jeall þeoſ mære ge sceafte :
17 Syllic wæſt fe rige beam ȝic ynnū fah for wounded
18 mid wommū ge seah ic wuldref treow . wædum geweor
19 ȝode wynnū scinan ge gyred mid golde gimmer hæfdon
20 be wrigene weorðlice wealdes treow : , hwaðore ic
21 þurh þær gold ongytan meahte earmra ærgewin
22 þær hit ærest ongan fwestan on þa swiðran healf
23 eallic wæſt mid jurgum gedrefed . Forht ic wæſt for
24 þære fægran ge yñhœ ge ȝeah ic þær fuſe beacen.

At the bottom of the interface, there are various navigation and zoom controls, including a zoom slider set at 34% and a scale indicator of 1:1.

Содержание

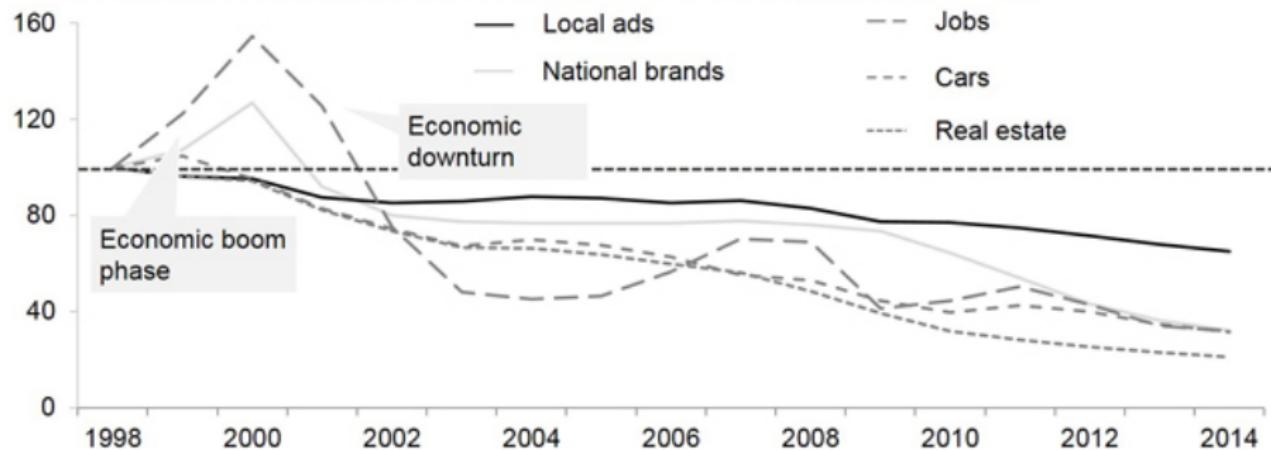
- 1 Оцифровка текстов: обзор
- 2 Оцифровка текстов: концептуальная сторона
- 3 Разметка

Как оцифровывать?

- Почему актуален вопрос «как оцифровывать?»
- Важными могут оказаться неожиданные аспекты

Если сохранять при оцифровке нужно только важное, то как решить, что важно?

Volumes for different categories of advertising in German newspapers (1998 = 100)



Что важно тут?



Традиционная текстология

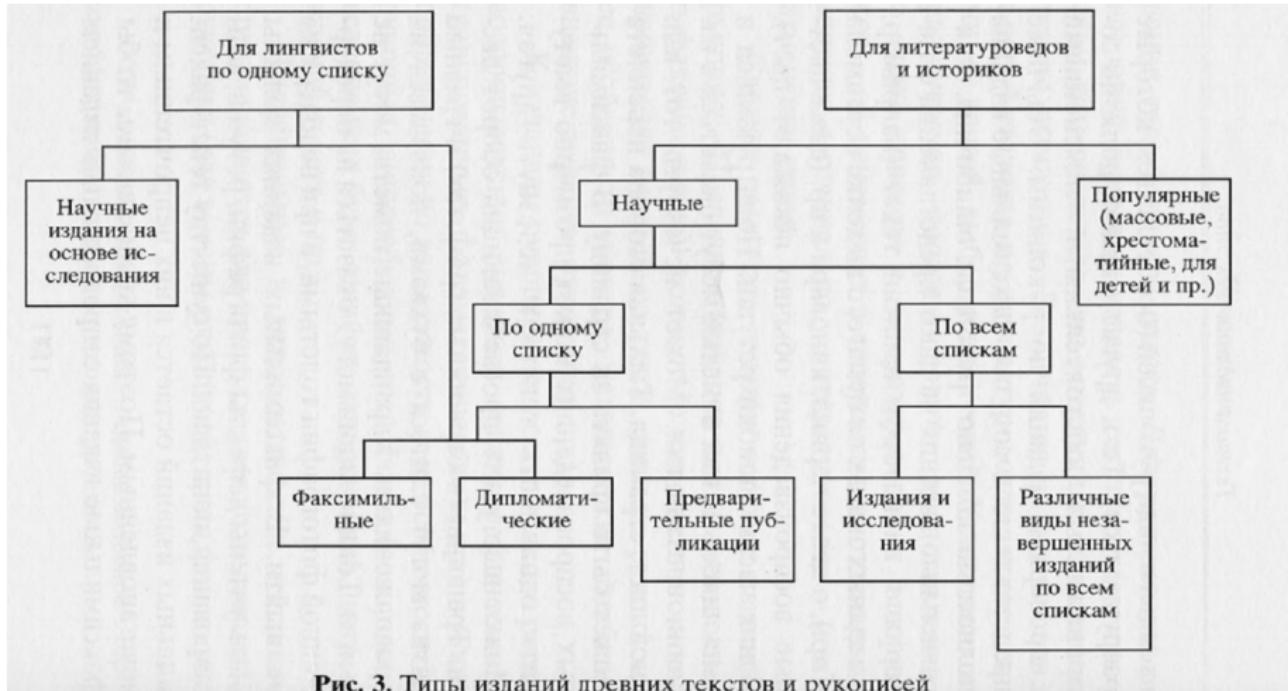


Рис. 3. Типы изданий древних текстов и рукописей

Дипломатические и критические издания

Herman Melville's Typee
A Fluid-Text Edition Upper Frame: Transcription ▾ P 17 ▾ Open Coordination ROTUNDA
Lower Frame: Base version ▾ P 17 ▾ Open

5 my was immediate flight; but in endeavoring to gain my
6 legs I fell back an{d} & rolled down a little grassy precipice
7 upon whose verge I had been laying; +— t +The shock slight
8 shock seemed to rally my faculties & starting to my
9 feet I fled wildly down the path I had just ascended.
10 The next moment without looking backward, I became
11 sensible from the feindish yells behind me that
12 As soon as ^{my enemies} the savages were in hot pursuit. Animated to
& heedless of & unheeding ^{the injury I had received tho'} the blood ^{stung}
13 madness by their fearful outcries & I now rushed down ^{flowing from my wound trickled}
over into
14 the mountain side with ^{a third of the distance} my eyes and almost blinded them
In a few short time I had descended nearly half-way & the my ^{almost}
15 ♫ & Suddenly a terrific howl burst upon my ear, & in ^{moment} the savages
16 the same & instant I instantly shrank to one side
17 from a fearful whizzing & noise in my ear — a heavy ^{noise}
little distance off & seemingly engaged in some
violent alteration respecting me. ¶ My first impulse
was immediate flight; but in endeavoring to gain my
legs I fell back & rolled down a little grassy precipice
upon whose verge I had been laying The slight
shock seemed to rally my faculties & starting to my
feet I fled wildly down the path I had just ascended.
The next moment without looking backward, I became
sensible from the feindish yells behind me that
my enemies were in hot pursuit. Animated to
madness by their fearful outcries & heedless of the injury I had received tho' the blood flowing from my wound trickled over
the mountain side with the speed of the wind
¶ In a short time I had descended nearly a third of the distance & the savages ceasing their cries, when Suddenly a terrific
same moment I instantly shrank to one side
from a fearful whizzing noise — a heavy

Текстологические задачи печатной эпохи

- Что считали важным текстологи? Можно ли на это опереться?
- Аван-текст: черновик/беловик/копия
- Первая публикация
- Прижизненные издания
- Разночтения и редакции
- Задача текстолога: восстановление текста наиболее близкого авторскому замыслу (принцип последней авторской воли)

- ... the text should be seen fundamentally as a function of the document. For, after all,... it is documents that we have, and documents only. (Gabler 1995)
- Основная идеологема — уход от принципа реконструкции воли автора к идее максимально точной передачи текста документа
- Экономические причины — удешевление дипломатических изданий (транскрипции), удешевление публикации больших архивов
- Технологические причины — электронная публикация вовлекает читателя и облегчает восприятие многослойного документа

Академический подход к цифровому изданию

The Walt Whitman Archive: The Whitman Archive's approach to editing is to establish documentary texts rather than to reconstruct what are assumed to be authorially intended texts.

The Walt Whitman Archive

Ed Folsom & Kenneth M. Price, Editors

SUPPORT THE ARCHIVE

Search the Archive Go

Published Works
Books by Whitman; Periodicals; Translations

In Whitman's Hand
Manuscripts; Notebooks; Scribal Documents

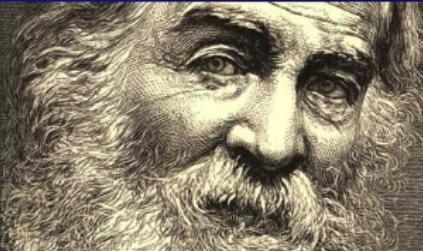
Life & Letters
Biography; Chronology; Letters

Commentary
Bibliography; Criticism; Disciples; Reviews

Pictures & Sound
Audio; Portraits of Whitman

Resources
Downloads; Manuscript Catalogs; Teaching; More...

About the Archive
Follow Us; Tour; People; Policies; More...

A detailed engraving portrait of Walt Whitman, showing him from the chest up, wearing a dark coat and a white collar. He has a full, bushy beard and hair.

News & Updates

May 2016
With support from the National Historical Publications and Records Commission, we've published 67 additional [letters](#) from the post-Reconstruction years.

April 2016
A new Whitman journalistic series has been discovered! "Manly Health and Training" was published in the *New York Atlas* in 1858. Find it in the [new double issue](#) of the *Walt Whitman Quarterly Review*.

December 2015
A new section of the Archive devoted to Whitman's [marginia and annotations](#) is now available. Thanks to the National Endowment for the Humanities for support of this work.

[Facebook](#) [Twitter](#) [YouTube](#) [More News](#)



Многослойный документ

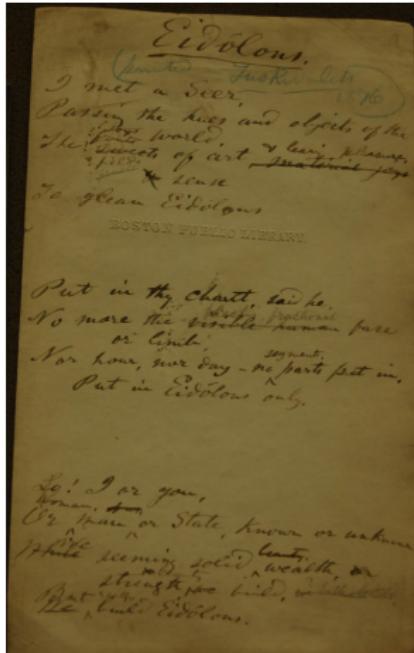


Рис.: Сканированное изображение

Eidolons.
(printed in Two Rivulets 1876)

I met a Seer,
Passing the hues and objects of the
The world.
The sweets of art & learning, pleasure,
& sense
To glean Eidolons

BOSTON PUBLIC LIBRARY.

Put in thy chant, said he.
No more the ~~visible~~ ^{fractional} human face
or limb!
Nor hour, nor day—no parts put in.
Put in Eidolons only.

Lo! I or you,
Or Woman ^{about three letters deleted, illegible} or State, known or unknown
We seeming solid ^{beauty} wealth or
strength we build, & ^{better} build;
But ^{we} build Eidolons.

Рис.: Транскрипция

Вместо «что является авторской волей?»
«что является значимым в документе?»

Архив Марианны Мур

<http://moorearchive.org/>

Moore Archive

EDITIONS AND RESOURCES

EDUCATION AND SCHOLARSHIP

ABOUT



Home / testing

Thumbs



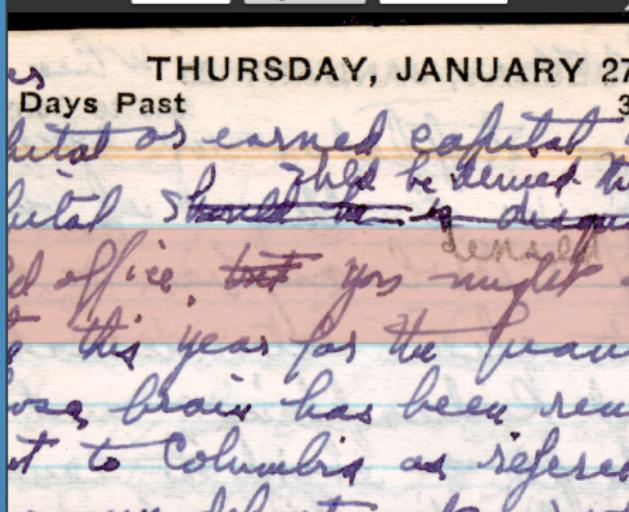
Image-Text



Show All Links

07-03-11

0021-verso



ever

inherited

capital or earned capital or used

shld be denied the right

capital should be is disqualified to

enses

hold office, but you might as well

vote this year for the peanut king
(whose brain has been removed &
lent to Columbia as referee in the polit
economy department—) who cannot
debate a matter for or against—who
just rules. Seriously—the country would
anyone



Что необходимо сохранить?

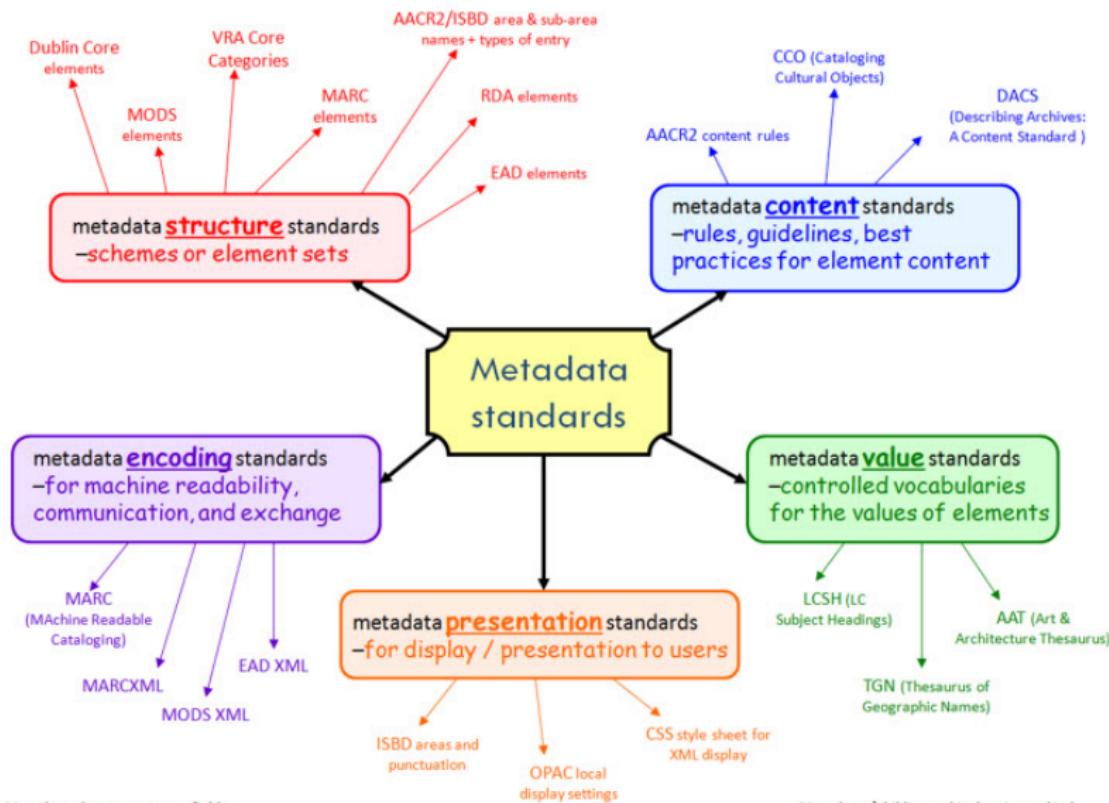
- Текст
 - Лингвистическое содержание
 - Особенности начертания: некоторые свойства текста могут быть переданы шрифтом.
- Материальный носитель и его особенности

Что можно добавить?

Специфически цифровое:

- Метаданные
- Исправления
- Конвертируемость в современный вид

Стандарты метаданных



Содержание

1 Оцифровка текстов: обзор

2 Оцифровка текстов: концептуальная сторона

3 Разметка

Что такое разметка?

Специальные символы, которые позволяют сделать обработку текста автоматической, указания для компьютера.

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="utf-8">
5     <title>Язык разметки</title>
6   </head>
7   <body>
8     <h1>Язык разметки<h1>
9     <p><strong>Язык разметки</strong> в компьютерной терминологии – набор
10    символов или последовательностей, вставляемых в текст для передачи
11    информации о его выводе или строении. Принадлежит классу
12    компьютерных языков.</p>
13    <h2>Структура языка</h2>
14    <p>Текстовый документ, написанный с использованием языка разметки,
15    содержит не только сам текст но и дополнительную информацию:</p>
16    <ul>
17      <li>указание на заголовки</li>
18      <li><mark>выделения</mark></li>
19      <li>списки</li>
20      <li>и т. д.</li>
21    <p>В более сложных случаях язык разметки позволяет вставлять в
22    документ <i>интерактивные элементы</i> и содержание других
23    документов.</p>
24  </body>
25 </html>
```

HTML

Язык разметки — HTML Academy

Структура языка

Текстовый документ, написанный с использованием языка разметки, содержит не только сам текст но и дополнительную информацию:

- указание на заголовки
- **выделения**
- списки
- и т. д.

В более сложных случаях язык разметки позволяет вставлять в документ *интерактивные элементы* и содержание других документов.

Основные теги:

- <cbml:panel> (n, characters)
- <cbml:balloon> (rendition, who, type)
- <cbml:caption>
- <floatingText> (subtype, type)
- <div> (type=panelGr, advert, etc)
- <sound>



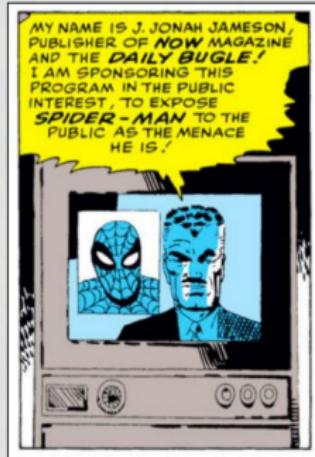
Примеры разметки:

```
<cbml:panel ana="#action-to-action"
    characters="#cap #anon_man" n="5" xml:id="eg_000">
    <cbml:caption> Cap acts quickly to tranquilize the gun-happy
pedestrian... </cbml:caption>
    <cbml:balloon type="speech" who="#cap" xml:id="eg_007">
A little <emph
    rendition="#b">sleep</emph> will do wonders for you!
</cbml:balloon>
    <sound> SPLAT! </sound>
    <cbml:balloon type="speech" who="#anon_man"> Ugh!
</cbml:balloon>
</cbml:panel>
```



Примеры разметки:

```
<cbml:panel characters="#spidey #jjj" n="3" xml:id="eg_ae1">
  <cbml:balloon rendition="#uc" type="audio" subtype="telecast"
  who="#jjj" xml:id="eg_006">
    My name is J. Jonah Jameson, publisher of <title
  rendition="#b">Now</title> magazine
    and the <title rendition="#b">Daily Bugle</title><emph
  rendition="#b">!</emph> I am
    sponsoring this program in the public interest, to expose
    <emph rendition="#b">Spider-Man</emph> to the public as the
  menace he is!
  </cbml:balloon>
</cbml:panel>
```



КАК МНОЖАТСЯ СТАНДАРТЫ:

(СМ.: ЗАРЯДНЫЕ УСТРОЙСТВА, КОДИРОВКИ, МГНОВЕННЫЕ СООБЩЕНИЯ И Т.Д.)

СИТУАЦИЯ:

ЕСТЬ 14
КОНКУРИРУЮЩИХ
СТАНДАРТОВ.

14?! АБСУРД! НАМ
НЕОБХОДИМО
РАЗРАБОТАТЬ ОДИН
УНИВЕРСАЛЬНЫЙ
СТАНДАРТ НА ВСЕ СЛУЧАИ
ЖИЗНИ.

ДА!



СКОРО

СИТУАЦИЯ:

ЕСТЬ 15
КОНКУРИРУЮЩИХ
СТАНДАРТОВ.

TEI: Text encoding initiative

- стандарт электронной публикации
- сообщество, вырабатывающее стандарт
- множество описаний
- появился в ноябре 1987



Что такое Text Encoding Initiative

- стандартизованный язык разметки на основе xml
- способ электронной публикации, при которой основные элементы документа/текста выделяются с помощью разметки и могут быть легко идентифицированы (поисковой системой, программой обработки и т.д.)
- способ представления документов и коллекций, обеспечивающий их взаимную совместимость (правильно структурированный документ xml)
- структура (framework) для стандартизованной разметки документа

- члены консорциума: 71 институт, 190 индивидуальных участников
- десятки воркшопов, школ и конференций в год
- Journal of the Text Encoding Initiative
- 1636 страниц гидлайнов в pdf
- примеры, образцы, мультиязычность, многоуровневый поиск, каталоги ресурсов

TEI: смысловые модули

- Все метаданные: о тексте, о воспроизведимом носителе, об электронном издании и его редакторах
- сегментация текста: структурные части документа — главы, титульный лист, эпиграф, сцены, строфы
- таблицы, формулы, графики, рисунки, музыкальные записи
- внутренние связи между элементами текста, ссылки
- графика и типографические элементы: особые буквы, знаки, старое написание
- критический аппарат: примечания, комментарии

Как устроен XML

Указание, что это документ XML, составленный в соответствии с инструкциями версии 1.0 в кодировке UTF-8

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<person xmlns="urn:mydomain:people"  
job="developer">
```

Элемент "person",
определенный в
пространстве имен

```
<firstname>John</firstname>  
<lastname>Doe</lastname>  
<birthday>1970-01-25</birthday>
```

Это все подэлементы "person",
которые находятся в том же
пространстве имен

```
</person>
```

Закрывает тэг "person"

XML example:

Recipe for Peanut Butter on a Spoon

```
<?xml ==?>version"1.0" encoding"UTF-8" "no" standalone=
<!DOCTYPE recipe PUBLIC "-//Happy-Monkey//DTD RecipeBook//EN
"http://www.happy-monkey.net/recipebook/recipebook.dtd">

<recipe>

<title>Peanutbutter On A Spoon</title>

<ingredientlist>
    <ingredient>Peanutbutter</ingredient>
</ingredientlist>

<preparation>Stick a spoon in a jar of peanutbutter, scoop
and pull out a big glob of peanutbutter.</preparation>

</recipe>
```

Работа с разметкой

werner_gebirgsarten_1787 (CN)

Bild: 0014 ▾

<< vorherige Seite

1:1

fit

63%

Kurze Klassifikation und Beschreibung

Ein Theil des Granits scheint das Grundeigtheit auszumachen. Der
Granit führt Metalle, besonders Zinn und Eisen.

§ 7.

Eine Granitarth, die eine befondere Gebirgsart zu seyn.
scheint.

Dorjenige Granit, welcher reich mit dem Glimmer geziest, heißt
fuer den Olimmer Granit. In diesem Glimmer ertheilt, steht von
seit seiner Erzeugung, als der eigentliche Granit zu seyn, und eine eigene
Gebirgsart anzusein. Solle sich hier noch nicht der Name bezeichnen, so
müsste man dieser Gebirgsart auch einen eigenen Namen geben, und
dann die jüngst Schriften nennen. Ich habe noch keinen Stangen-<lb/>
Kohle-<lb/> mit dieser Gebirgsart gefunden, im u. #0364; bringen ko. #0364;nmitte f. mit dem
eigentlichen Gra-<lb/>

nite (§. 6.) ziemlich überzeugt. Diese Bergart fu. #0364; hrt gleichfalls

Metalle. <lb/>

Sie ko. #0364; met mit wenigen Olimmer zu <placeName>Scharfenberg</placeName>,
<placeName>Alten</placeName> und <placeName>Militz</placeName>, unweit

bergs</placeName>, ganz ohne Olimmer aber zu <placeName>Militz</placeName>, unweit

<placeName>Meißen</placeName>, und befor-<lb/>

ders zu <placeName>Prieznitz</placeName>, und zu

<placeName>Frohnsdorf</placeName>, unweit <placeName>Dresden</placeName>,

desgleichen auch in der <placeName>überlaufite</placeName> vor.</p>

§ 8.

a.) Gneiss.

Der Gneiss besteht ebenfalls aus Gletschern, d) Quarz und Olimmer,
die aber in einem diebstoffen oder sonstigen Gestein mit einander ver-

verbunden

o Die Mineralogen übereinstimmen hier mit Jahr 1775, den Gletschern in
den Erklärunghen, die von dem Gneiss aus, dass er ein Gestein ist, aus
deren Natur, dass er einen Granit, oder einen Granit, oder einen
Gneiss, und auch andere vertheilten. Aber in das Gespräch der Dritte
findet Generat, so wie auch die meisten Autoren auch durch die wenig
genug beobachtet, und alle 4 Theile dieses Gesteins sind einzeln. Sie
sind im Jahr 1775, den Gletschern, das erste auf diese Weise

beschrieben.

Berlin 598-PX, 3 in: M 3505

offene Tickets: 2 (1 ganzes Buch)
Stand: Sat Mar 16 10:12:47 2013

Text Text/Bild ⚡ 8
0 - 9 - 27 1 - 9 - 26

9 <p> facse="#f0014" n="0"/>
10 <fw place="top" type="header">Kurze Klassifikation und Beschreibung</fw><lb/>
11 Ein Theil des Granits scheint das Grundeigtheit auszumachen.
12 Granit fu. #0364; hrt Metalle, besonders Zinn und Eisen.</p><lb/>
13 <div ns="g">
14 <head>§. 7.<lb/>
15 </head>
16 <head>Eine Granitarth, die eine befondere Gebirgsart zu seyn</lb/>
17 <head>Gneiss</lb/>
18 <p>Derjenige Granit, welcher theils mit dem Glimmer zugleich, theils<lb/>
19 aus dem Glimmers Hornblende in feinem Gemenge entha. #0364; hrt. Gneiss</lb/>
20 wird neuerer Erzeugung, als der eigentliche Granit zu seyn, und eine eigene<lb/>
21 der Gneissart ist eine wahre. Solche Gneiss sind durch noch nicht best. #0364; ligent. <lb/>
22 mus#0364; gte man dieser Gebirgsart auch einen eigenen Namen geben, und<lb/>
23 ko. #0364; mte die f. #0364; gleich <hi rendition="#ff">BrauŬntstein</hi>
24 nennen. Ich habe noch keinen Stangen-<lb/>
25 <chot#0364;rl in ihr gefunden; im u. #0364; bringen ko. #0364;mmt die mit dem
26 eigentlichen Gra-<lb/>
27 nite (§. 6.) ziemlich u. #0364;berren. Diese Bergart fu. #0364; hrt gleichfalls
28 Metalle. <lb/>
29 Sie ko. #0364; met mit wenigen Olimmer zu <placeName>Scharfenberg</placeName>,
30 <placeName>Alten</placeName> und <placeName>Militz</placeName>, unweit
31 bergs</placeName>, ganz ohne Olimmer aber zu <placeName>Militz</placeName>, unweit
32 <placeName>Meißen</placeName>, und befor-<lb/>
33 ders zu <placeName>Prieznitz</placeName>, und zu
34 <placeName>Frohnsdorf</placeName>, unweit <placeName>Dresden</placeName>,
35 desgleichen auch in der <placeName>überlaufite</placeName> vor.</p>

51 </div>

Editor
undo redo
Zeichen
langes s runderst UHFFC n-dash
n-dash n-dash
Elemente einfugen
Milestones:
<Line> <lb/>
<orgName>
Eigennamen:
<placeName> <name>
Druckfehler:
<choice>sic=>corr</choice>
Normalisierung:
<choice>orig=>reg</choice>
Abkürzung:
<choice>abbr=>expans</choice>
Hervorhebungen:
#ub #g #s #E
#In #blue #red
Verknüpfungen:
#next #nest #xal:id
#prev #prev #xal:id

Suche
DOC
grep +
Anmerkung zu dieser Seite anlegen

Краудсорсинг



UCL Home > Transcribe Bentham > Transcription Desk

Justintona My Talk My Preferences My Watchlist My Contributions Log Out

Navigation

Edit

Discussion Watch History Protect Move Delete Refresh my messages

- Transcription Desk
- Blog
- Discussion Forum
- Recent changes
- Random page
- Contact Us
- Help

Search

Go Search

Toolbox

- What links here
- Related changes
- Upload file
- Special pages

Screencast-O-Matic.com

Editing JB/088/002/003

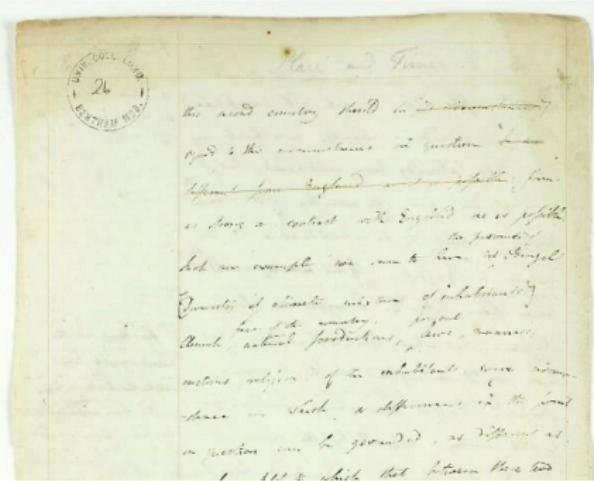
J P [+/-] ? ... Ab sic fr & - <->

[!File:088_002_003.jpg!520px[right]]
!!-- ENTER TRANSCRIPTION BELOW THIS LINE -->

this second country should in [del]

circumstances<lb>/>
regard to the circumstances in question be
as<lb/>
different from England as it is possible
form<lb/>/>
as strong a contrast with England as is
possible.<lb/>

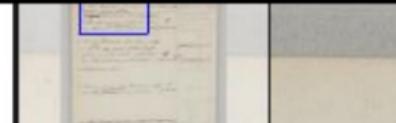
<!-- DO NOT EDIT BELOW THIS LINE -->
(Metadata:{(PAGENAME)})



Краудсорсинг

```
'''{{fullurl:JB/001/003/001|action=edit}} Click Here To Edit'''
<!-- ENTER TRANSCRIPTION BELOW THIS LINE -->
<note>Abstract</note>
<p>II. Hypothetical Profit, <add>Type larger <hi rend="underline">on the same score</hi></add> by Sale of an equal quantity
```

II. Hypothetical Profit, Type larger on the same score by Sale of an equal quantity 7



Type larger on the same score
Hypothetical Profit, by Sale of an equal quantity

С помощью разметки можно выделять не только структурные, но и семантические элементы тексты и делать их понятным компьютеру, увязывать между собой в разных текстах.

Это позволяет говорить о семантическом издании текста.

Transkribus

Transkribus v6.5.9 (27.03.2015, 15:08), Loaded doc: HTR Reichsgericht v2, ID: 301, Page 1, file: ZS-I-1902-198 (2).jpg*

Documents Structure Tools Transkribe Pages

Type Text Structure Read... O

Text graph

Text style

Page type

Links

Structure Type

Tags

Text

Live board Region H 4 2 / 2 H A I T R suggestions

1 ist vom Kläger bezahlt worden Anfang 1900 erwirkten Beklagter da
2 Zusatz Patent No . 10 245 zu dem Hagen Patent . 82785 vgl Blatt 15
3 und bat dem Kläger unter näheren Bedingungen vgl Blatt 20 Es eine Als
4 änderung des Vertrages vom März 1899 an , wonach Kläger an Stelle
5 des behaupteten es das Zusatz Patent benutzen sollte . Am 2 Februar
6 1900 fand hierüber bei einer zu fülligen Berechnung in Berlin ein
7 Ansprüche zwischen den Parteien statt , wobei Kläger dem Beklagten
8 Vorwürfe machte , daß er selbst setzt das Patent 82785 für werde
9 erklärt und chen bezüglich der Auslegung des Zusatz Patents um mehr

The screenshot shows the Transkribus application window. At the top, it displays the version (v6.5.9), date (27.03.2015, 15:08), and document information (HTR Reichsgericht v2, ID: 301, Page 1, file: ZS-I-1902-198 (2).jpg*). The interface includes tabs for Documents, Structure, Tools, Transkribe, and Pages. On the left, there's a sidebar with a tree view of the document structure and a list of terms (Type) such as 'Text', 'Structure', and 'Read...'. The main area contains the scanned document page with handwritten text in brown ink on a white background. Below the page, the transcribed text is displayed in a light brown box. To the right of the page, there are several toolbars and panels for editing and styling, including 'Text graph' (with icons for different text types like heading, list-item, drop-capital, etc.), 'Text style' (with options for font family, size, color, and background), and a 'Page type' dropdown. The bottom of the window features navigation icons for the document.

Остальные аспекты оцифровки

