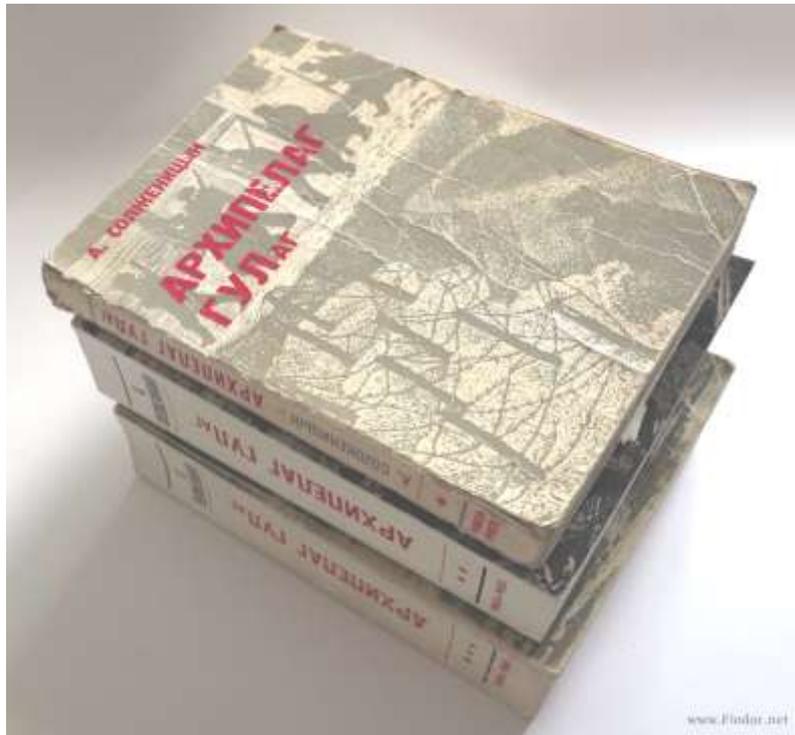


**ОПЫТ РАЗРАБОТКИ СТАНДАРТНОГО HTML-ФОРМАТА
ДЛЯ ОЦИФРОВКИ ИСТОЧНИКОВ СОВЕТСКОГО ПЕРИОДА**

Григорий Белонучкин
Центр «Панорама»

istnet.org

Источники



ОПЫТ ХУДОЖЕСТВЕННОГО ИССЛЕДОВАНИЯ



ОФИЦИОЗ ВО ВСЁМ ЕГО МНОГООБРАЗИИ

Цель

- Стандарт – не директивный документ, а набор приёмов, которые создатель любого интернет-ресурса волен применять полностью или частично
- Адресован энтузиастам-одиночкам, которые только и занимаются оцифровкой советского наследия

Параметры стандарта

ДАНО (тексты XX века)

- Современный алфавит
- Эталонный печатный оригинал, массовый тираж
- Традиционная библиогр. ссылка (изд., т., стр.)
- Нехитрое шрифтовое оформление
- Мало иллюстраций, формул, диакритики

ТРЕБУЕТСЯ

- Библиографическая ссылка ↔ гиперссылка
- User-friendly URL (вплоть до адреса конкр.страницы)
- Совместимость с основными поисковиками
- Простая техника, не требующая программистских навыков

Что не устраивает:

- 1) Библиотечные оболочки – закрытый авторский программный код; защита от копирования; отсутствие постоянного URLa у книги, а тем более – у страницы
- 2) Графические форматы – обычно недораспознаемость; смешение собственно текста со служебной информацией; неконвертируемость в Word
- 3) PlainText – теряется шрифтовое выделение и иерархия заголовков
- 4) HTML – номера страниц и примечания разрывают контекст **и это единственная проблема формата HTML**

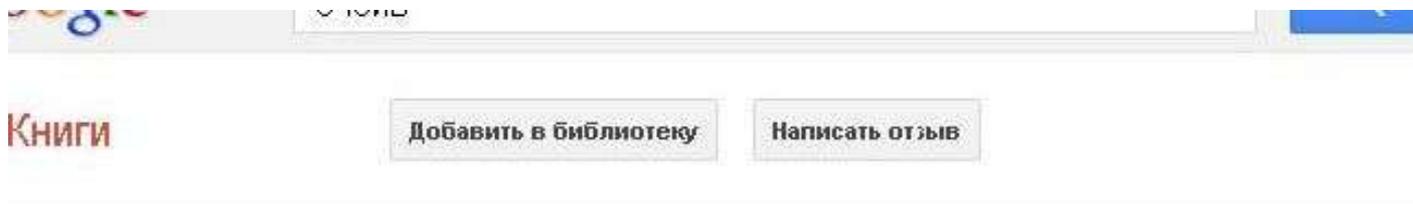
Автораспознавание-1

The image shows a screenshot of a document viewer interface. At the top, there is a header with a logo on the left and the text "ОНЛАЙН ПРОСМОТР ДОКУМЕНТА" and "Документ: Ленин (28.85 Mb)". To the right of the header are navigation controls, including a page number "21 / 53". Below the header, there are two tabs: "Эскизы" and "Поиск". Under the "Поиск" tab, there is a search input field containing the word "Октябрь" and a "Найти" button. Below the search field, it says "Найдено: 8". There are three search results listed:

- Стр. 10. енародного сердца с Октябрьской грозой, Пусть на полке Тургене...
- Стр. 15. Октябрьские рассветки п сумерки С ледовиты...
- Стр. 21. Октябрь — месяц проспш, листопада, Треско...

A red arrow points to the third search result. To the right of the search results is a large document page. The page is mostly blank with some faint text. A green arrow points to a line of text on the page: "Октябрь — месяц просши, листопада, Тресковой солки и рябиновых бус...".

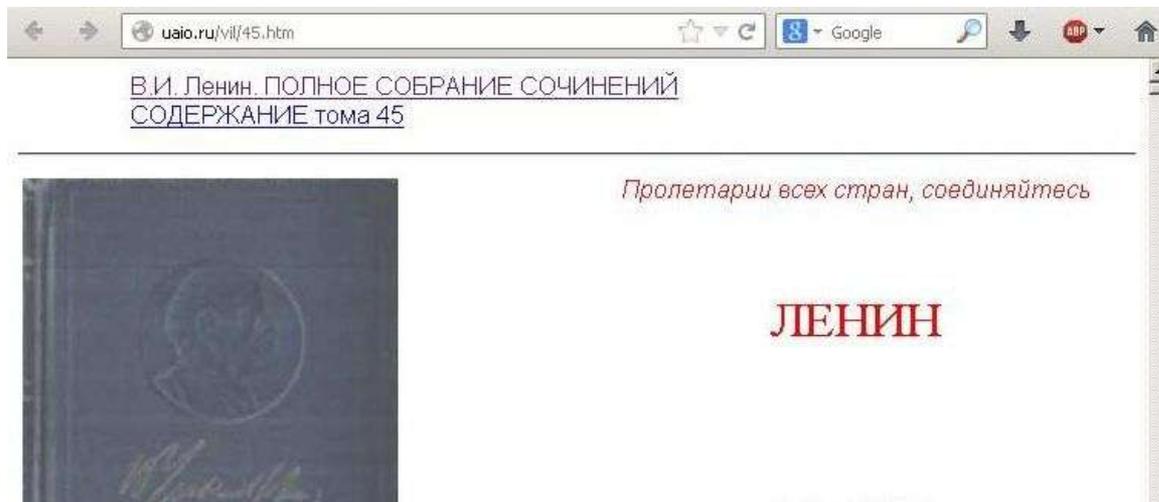
Автораспознавание-2



Часто встречающиеся слова и выражения

августа альманах апреля аудитории большой БРИК Бурлюк вериулся |
Владимир Маяковский Война вопрос воспоминаниям времени
выступление вышел ГАЗЕТЕ говорил года ГОСИЗДАТ декабря
других жизни журнале зале **знаю** иад иапечатаиа статья иапечк
иас иачал иашей иесколько иет Известия издание иитерес **иовой**

HTML: разрыв цитаты



В условиях Советской власти, такой капитализм, который сознательно допускается и ограничивается рабочим классом. От государственного капитализма стран, имеющих буржуазные правительства, наш государственный капитализм отличается весьма сущест-

297

РУССКОЙ КОЛОНИИ В СЕВЕРНОЙ АМЕРИКЕ

венно, именно тем, что государство у нас представлено не буржуазией, а пролетариатом, который сумел

Цитата не ищется

"наш государственный капитализм отличается весьма существенно"

на сайте: uaio.ru [расширенный поиск](#)

Точная цитата "наш государственный капитализм отличается весьма существенно" в кавычках нигде не встречается. Показаны результаты по запросу без кавычек. [?](#)

Область поиска: сайт — [uaio.ru](#)

[Разместит](#)

[«наш...»](#)

[Курс политической экономии. Том I. Досоциалистические способы...](#)

Процесс формирования монополярной цены **существенно отличается** от образования рыночных цен в условиях домонополистического ... Между тем **государственно-монополистическое** регулирование направлено прежде всего на сохранение самих устоев **капитализма, а весьма...**

[bse.uaio.ru > 73/02.htm](#)

[Курс политической экономии. Том I. Досоциалистические способы...](#)

Государственно-монополистический капитализм не является ни особой стадией, ни особой фазой ... Это различие **весьма существенно**, и его ... Например, экономические категории, выражающие признаки и свойства **капитализма, существенно отличаются** от категорий...

[bse.uaio.ru > 73/01.htm](#)

[Курс политической экономии. Том II. Социализм \(ч.1\)](#)

Вместе с тем она **существенно отличается** от **государственной** формы хозяйствования, где средства производства и труд обобществлены, как уже ... При **капитализме**, в эпоху свободной конкуренции, экономическая роль буржуазного государства была **весьма** ограниченной.

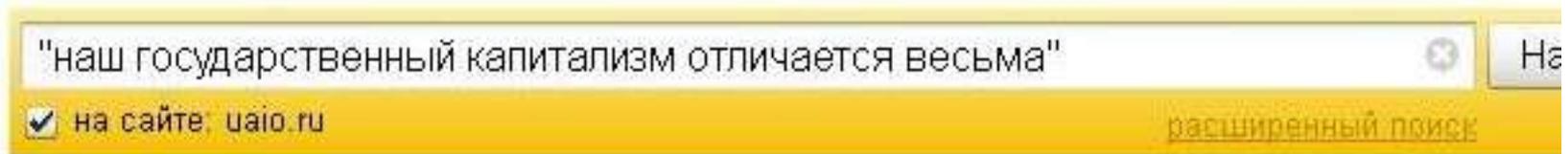
[bse.uaio.ru > 73/03.htm](#)

[Ленин ПСС издание 5 том 45](#)

Чрезвычайно **существенно** поэтому, чтобы в дополнение к работе соответствующих **государственных** учреждений, в ... От **государственного капитализма** стран, имеющих буржуазные правительства, **наш государственный капитализм отличается весьма** существ

[uaio.ru > vil/45.htm](#)

Мнимая точка



Область поиска: сайт — uaio.ru

Ленин ПСС издание 5 том 45

От государственного капитализма стран, имеющих буржуазные правительства, **наш государственный капитализм отличается весьма** существ. 297 русской колонии в северной америке.

[uaio.ru > vil/45.htm](http://uaio.ru/vil/45.htm)



Критерии оптимальной разметки

СПЕЦИФИКАЦИИ HTML

- 4.01, 5.1, xhtml
- Включают теги, не поддерживаемые ни одним браузером
- Строгие критерии валидности кода, необязательные для браузеров

НЕ КРИТЕРИЙ

БРАУЗЕР-СОВМЕСТИМОСТЬ

- IE
- Firefox
- Opera
- Chrome
- Safari-iPad

ЖЕЛАТЕЛЬНЫЙ КРИТЕРИЙ

ПОИСК-СОВМЕСТИМОСТЬ

- Яндекс
- Google

ОБЯЗАТЕЛЬНЫЙ КРИТЕРИЙ

Человеко-понятный URL

ТРАДИЦИОННЫЙ СПОСОБ (через PHP)

<http://site.ru/index.php?book=pushkin&vol=2&p=16>



<http://site.ru/pushkin/2/16>

ПРЕДЛАГАЕМЫЙ СПОСОБ (в пределах HTML)

<http://pushkin.site.ru/2/index.html#16>



<http://pushkin.site.ru/2/#16>

<http://пушкин.сайт.рф/2/#16>

Буквы (lat,кир) – только до слэша

Л.И.Брежнев, Соч., т.1, с.4

brezhnev.su/1/#4

2-й текст 2-го тома Л.И.Брежнева:

brezhnev.su/2/#02

Ведомости ВС СССР, 1991, №52, ст.1531

vedomosti.sssr.su/1991/52/#1531

XXV съезд КПСС. Стеногр. отчет, т.1, с.322

25съезд.кпсс.рф/1/#322

Метка номера страницы (главы, строки, строфы etc.)

- `...` ~~Яндекс~~ Google
- `...` ~~Яндекс~~ Google
- `...` ~~Яндекс~~ ~~Google~~
- `...` Яндекс Google

Выделение границы страниц

- FONT COLOR
- SPAN STYLE
- U.../U
- FONT STYLE

Яндекс Google

~~Яндекс Google~~

Яндекс Google

Яндекс Google

Номер страницы во всплывающей подсказке

- A TITLE
- SPAN TITLE
- DIV TITLE
- FONT TITLE

~~Яндекс~~ Google

~~Яндекс~~ Google

~~Яндекс~~ Google

Яндекс Google

Как это выглядит в браузере

Home Index Contents Search Glossary Help First Previous Next Last Up Copyright Author

- у нас **обнаруживались** те или иные
(FONT COLOR)
- у нас **вскрывались** те или иные
(U.../U...FONT STYLE="text-decoration:overline[...])
- у нас **возникали** те или иные 48 ← **всплывающий № стр.**
(FONT STYLE="background-color[...])
- у нас **выявлялись** те или иные
(FONT STYLE="border[...] для последнего и первого полуслова страницы)
- у нас **всплывали** те или иные
(FONT STYLE="border[...] только для первого полуслова)

Таблица стилей в HEAD

```
<style type="text/css">  
.ps {border: 1px; border-color: orange;  
border-style: solid none none solid}  
</style>
```

(Стиль PageStart: граница – 1 пиксель;
оранжевая; сверху есть, справа нет, снизу
нет, слева есть)

Разметка на стыке страниц

марксизма-ленинизма,

пролетар

ского

интернационализма,

(Бурные аплодисменты.) 9
движения, добиваться преодоления труднос
изма, пролетарского интернационализма,
ещания марксистско-ленинских партий всего

Трудоёмкость

Сводится к пяти операциям:

- Разбивка на абзацы: `</p><p>`
- Разметка заголовков трёх-четырёх уровней: `<h1>...</h1>` etc.
- Простановка кодов на границах страниц
- Разметка жирных, курсивных, латинских с диакритикой и других нестандартных контекстов
- Вставка типовых «шапки» и «хвоста» в файл

В сравнении со сплошной вычиткой (а она необходима!) эта работа – сущие пустяки