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

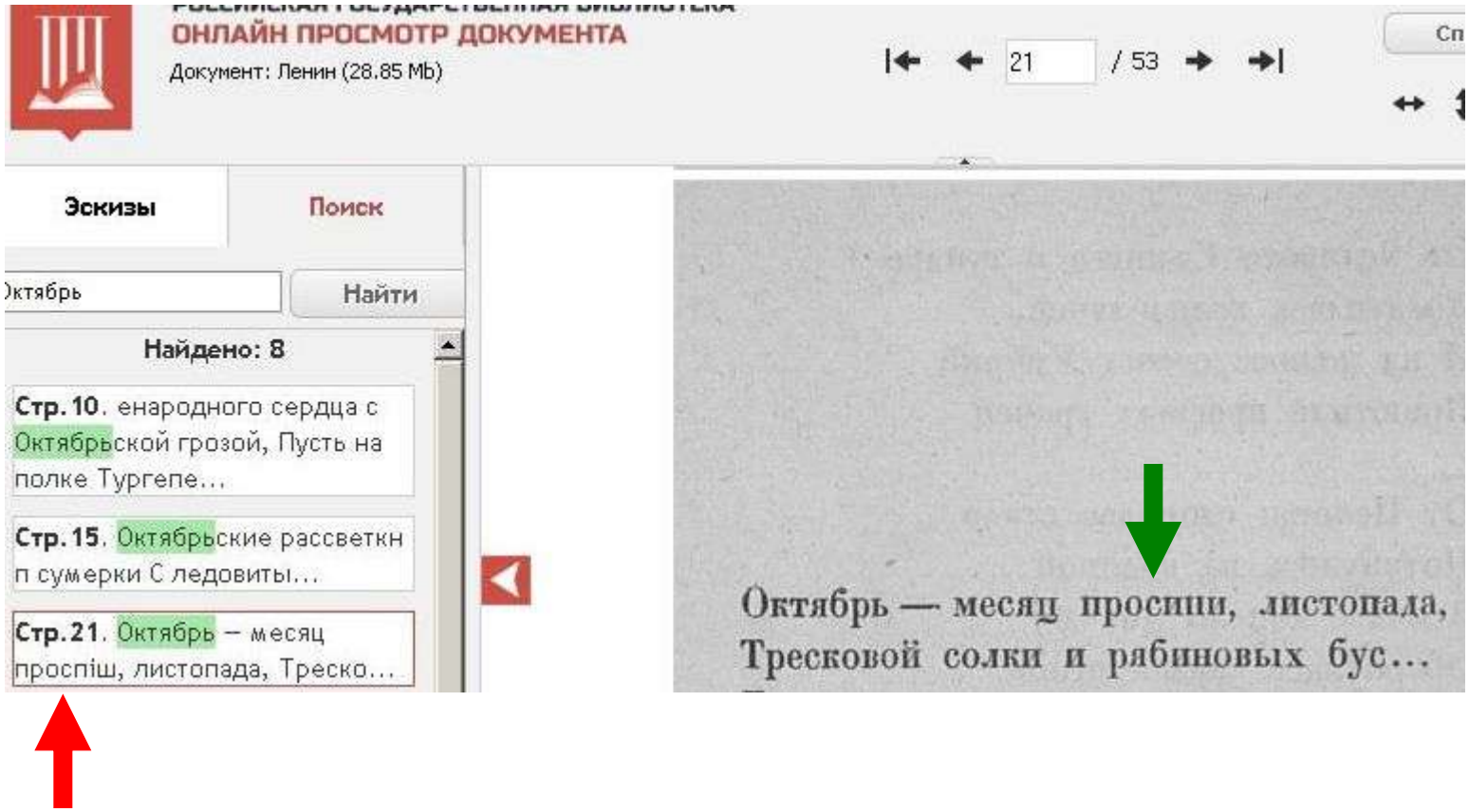
ТРЕБУЕТСЯ

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

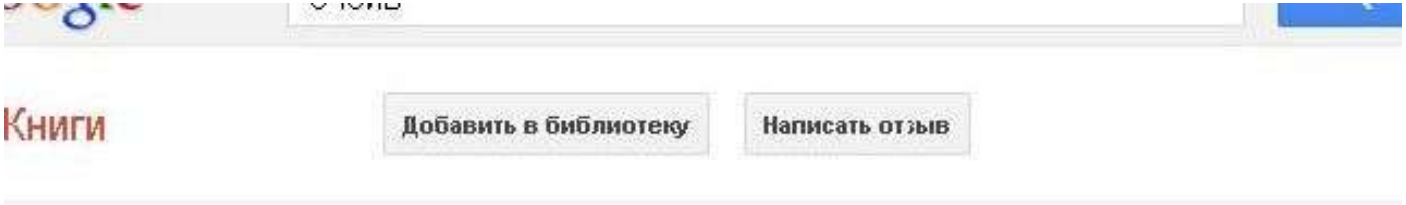
Что не устраивает:

1) Библиотечные оболочки – закрытый авторский программный код; защита от копирования; отсутствие постоянного URLa у книги, а тем более – у страницы
2) Графические форматы – обычно недораспознанность; 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

Авторасознавание-1



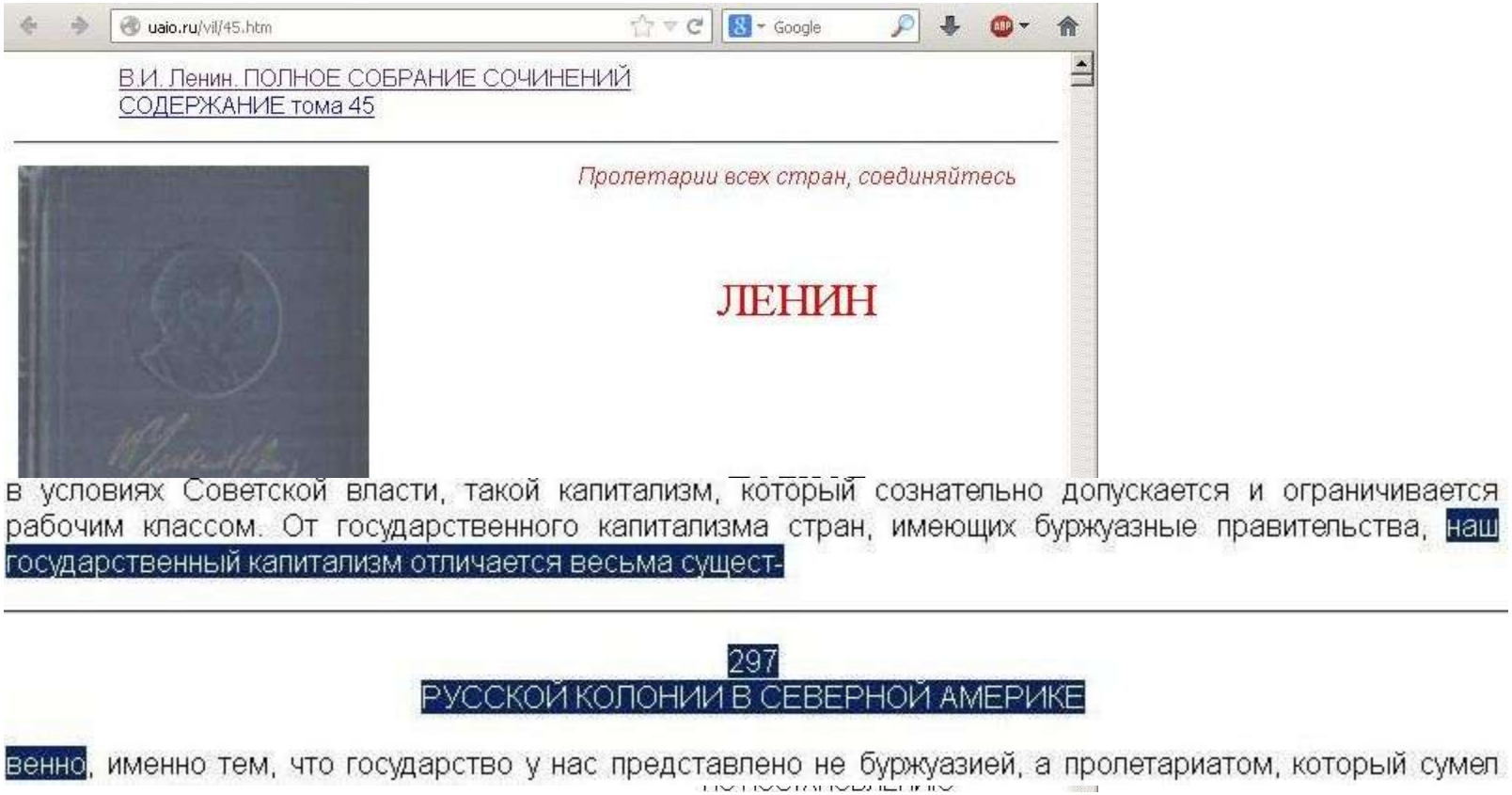
Авторасознавание-2



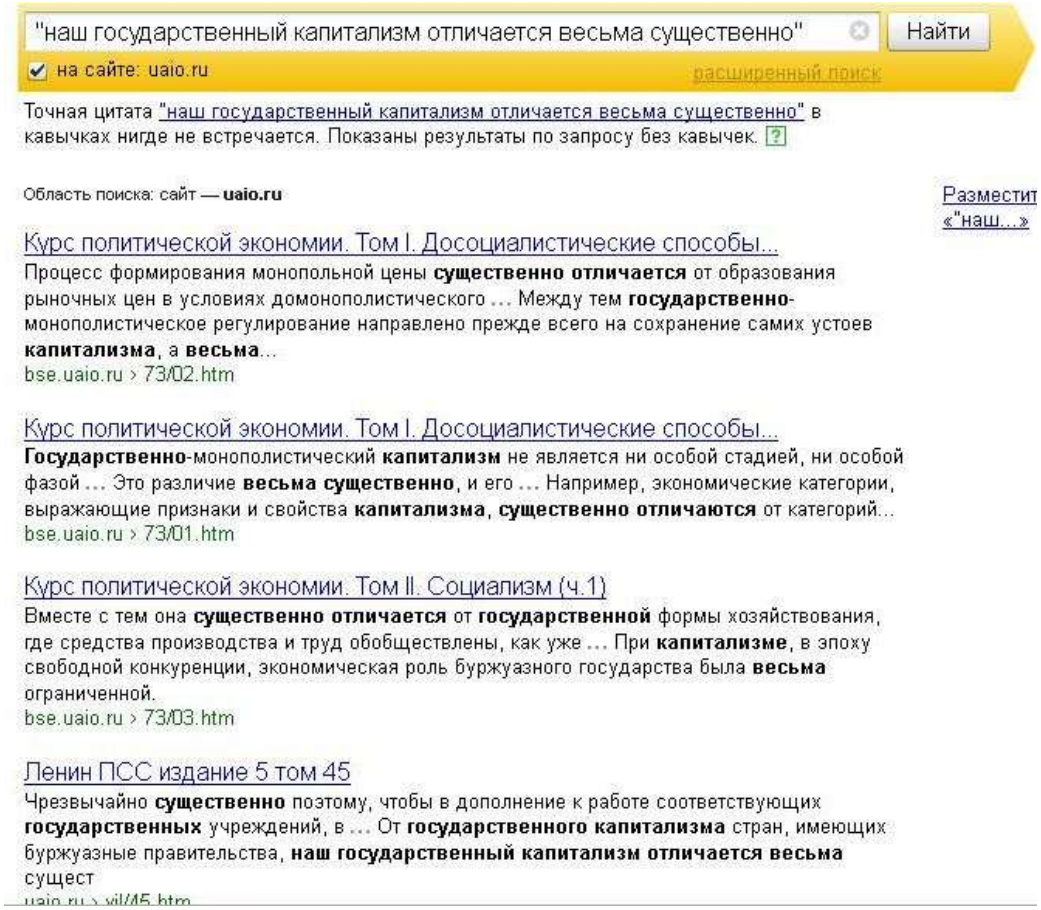
Часто встречающиеся слова и выражения

августа альмаиах апреля аудитории большой Брик Бурлюк вериулся
Владимир Маяковский Войиа вопрос воспомииаииям времеии
выступлеиие вышел газете говорил года Госиздат декабря
других жизии журиале зале зиаю иад иапечатаиа статья иапеча
иас иачал иашей иесколько иет Известия издаиие иитерес иовой

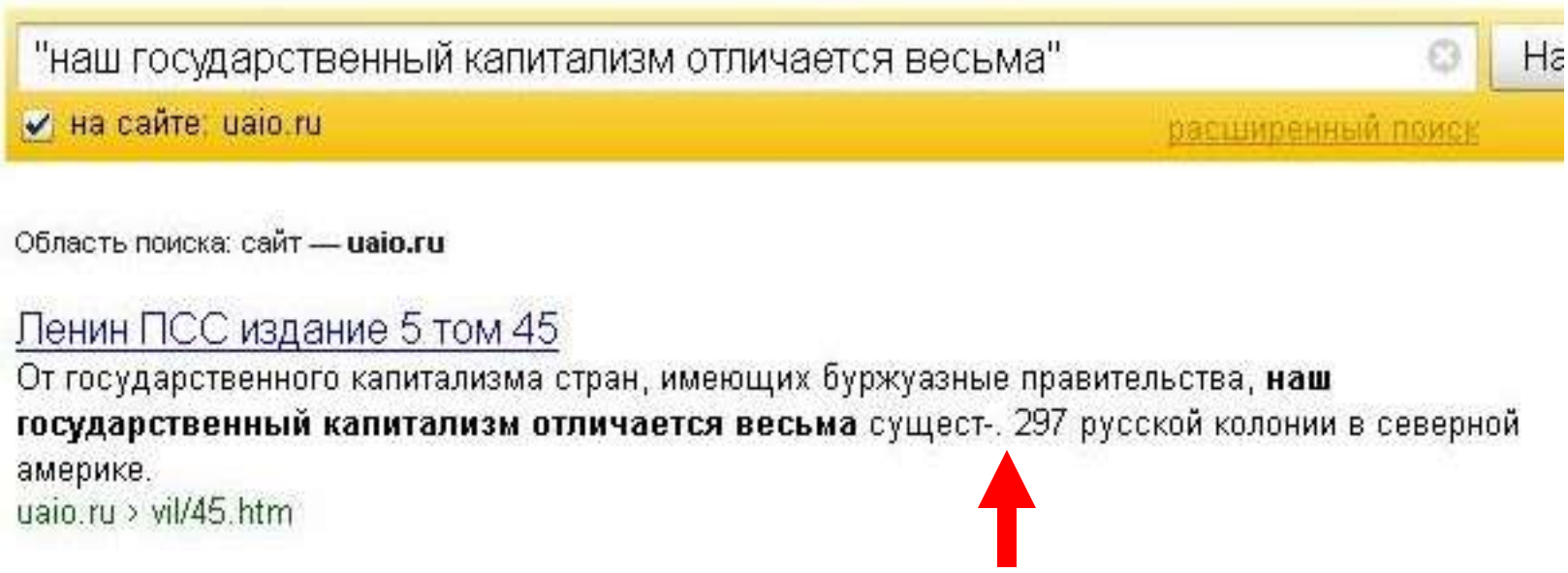
HTML: разрыв цитаты



Цитата не ищется



Мнимая точка



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.)

- <A NAME="1">…</A> ~~Яндекс~~ Google
- <A ID="1">…</A> ~~Яндекс~~ Google
- <SPAN ID="1">…</SPAN> ~~Яндекс~~ ~~Google~~
- <FONT ID="1">…</FONT> Яндекс Google

Выделение границы страниц

- FONT COLOR — Яндекс Google
- SPAN STYLE — Яндекс ~~Google~~
- U…/U — Яндекс Google
- FONT STYLE — Яндекс Google

Номер страницы во всплывающей подсказке

- A TITLE — ~~Яндекс~~ Google
- SPAN TITLE — Яндекс ~~Google~~
- DIV TITLE — ~~Яндекс~~ ~~Google~~
- FONT TITLE — Яндекс Google

Как это выглядит в браузере

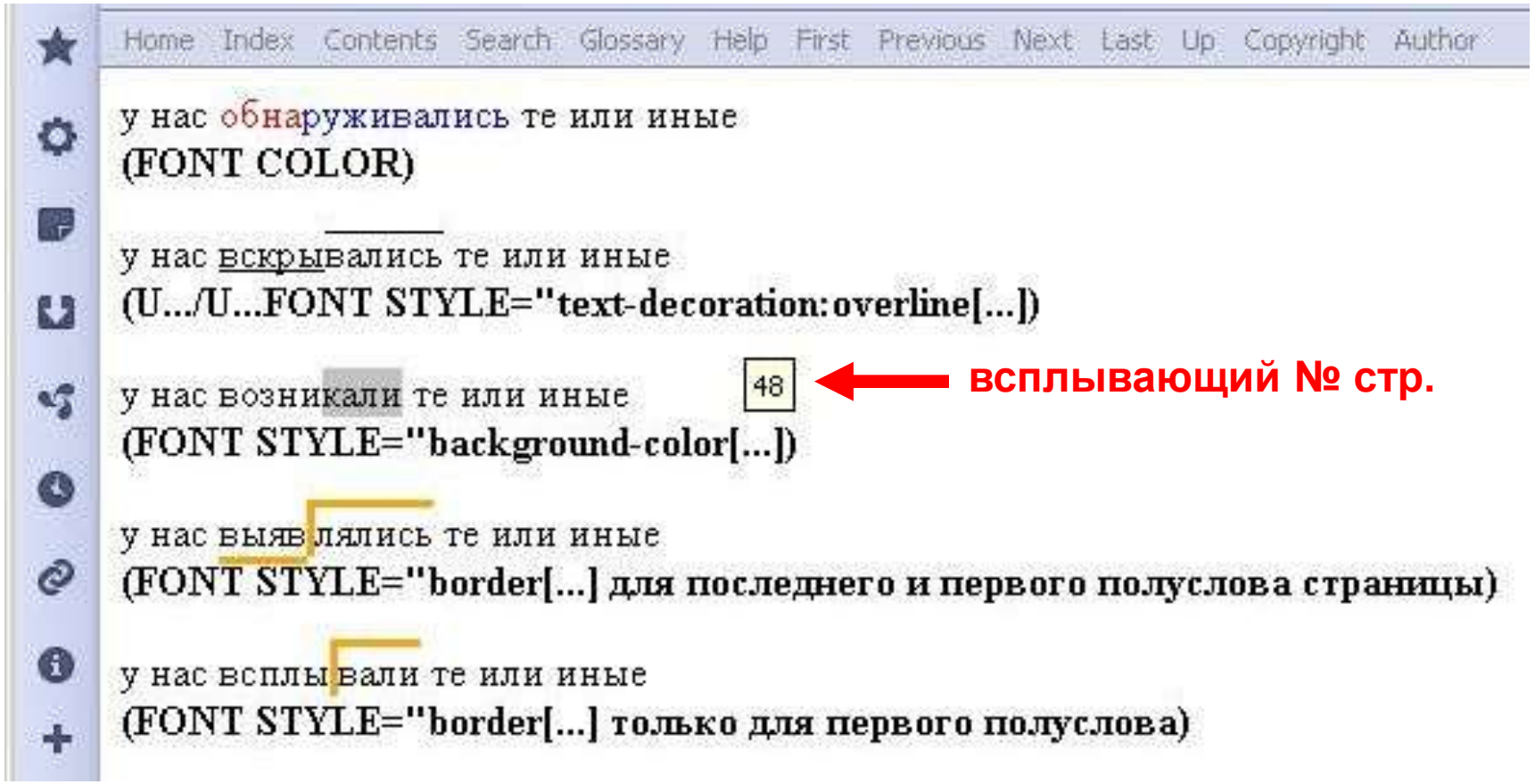


Таблица стилей в HEAD

```
<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)

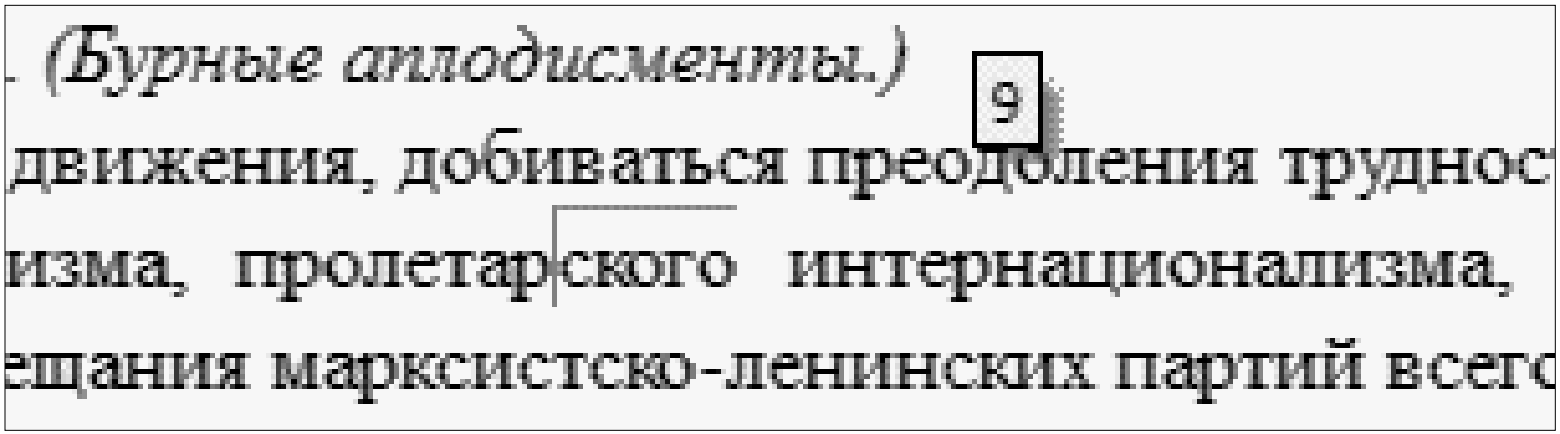
Разметка на стыке страниц

марксизма-ленинизма,

пролетар<font id="9" title="9"

class="ps">ского<font title="9">

интернационализма,



Трудоёмкость

Сводится к пяти операциям:

- Разбивка на абзацы:
- 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