Опыт разработки стандартного формата на основе html для оцифровки источников советского периода

© Г.В. Белонучкин
Информационно-исследовательский центр «Панорама»,
Москва
belonuchkin@gmail.com

Аннотация

В настоящем кратком сообщении рассматривается необходимость и возможность выработки стандарта представления текстов исторических источников XX века на русском языке в Интернете для удобства поиска, копирования и научного цитирования, а также излагаются результаты проведённого эксперимента по выработке оптимального варианта HTML-разметки, совместимого со всеми распространёнными браузерами и наиболее используемыми поисковыми системами.

1 Необходимость систематизации советских источников

Знание о советской истории, вероятно, будет неполным без прочтения «Архипелага ГУЛАГ» [16] или книг А. Некрича [6] и Р. Пихои [14], но, с другой стороны, невозможно составить адекватное представление о советском времени лишь по «Архипелагу...», не пролистав подшивку газеты «Правда» [15] и материалы съездов КПСС, не усвоив иерархию письменных источников, начиная с Полного собрания сочинений В. И. Ленина [11], занимавшего в этой иерархии место Библии [3].

Многие книги с «верхних ступенек» этой иерархии давно оцифрованы самодеятельными энтузиастами, но в разных форматах, зачастую неудобных для поиска (часто – только графических) и с бесчисленными опечатками. В Сети уже есть Собрание сочинений К. Маркса и Ф. Энгельса в качественном текстовом PDF [13]; многократно оцифрованы труды В. И. Ленина (особо стоит упомянуть PlainHTML-версию на сайте Василия Грозина) [12]; распознанные наполовину многотомники Л. И. Брежнева и Н. С. Хрущёва можно найти в электронной библиотеке «Нефть и газ» [2]; стенотчёты ряда съездов КПСС и пленумов ЦК – в Военном разделе Библиотеки Максима Мошкова [5]; есть в разных уголках Интернета ежегодные «Справочники партийного работника», отдельные тома «Документов внешней политики СССР». Довольно часто оцифрованные книги не выставляются в сети, а помещаются на файлообменные сайты (подчас в архивированном виде). Отыскать их трудно, а найти в них некий документ, даже по дословной закавыченной цитате, порой нельзя, потому что цитата оказывается «разорванной» посередине какой-нибудь служебной информацией – например, номером страницы или сноской.

Когда за дело создания корпуса оцифрованных книг берётся серьёзная организация (библиотека, фонд, коммерческий портал), результат нередко плачевен. Нанятый заказчиком программист пишет свою программную оболочку, в которой каждую страницу можно увеличить, уменьшить, растянуть и повернуть на 45 градусов, но ***найти*** её через общепринятые «ворота» Интернета – поисковые системы Яндекс и Google – проблематично; адрес страницы генерируется заново для каждого сеанса (т.е. невозможно создать извне постоянную ссылку на конкретный текст, тем более – на конкретный участок текста), а техническая поддержка требует постоянного личного участия программиста.

Это объяснимо: те, кто вложил деньги и труд в оцифровку источника, не хотят, чтоб результат их труда «воровали». Наш подход другой: никакой технической защиты, максимальное удобство для копирования, работы с текстом и распечатки.

Книги советского времени имели лишь ту функциональность, которую заложил в них издатель, и наделять их чем-то бóльшим вряд ли стоит. Гиперссылки с одного источника на другой уместны в интернет-версии преимущественно там, где эти ссылки сделаны самим автором. Произвольное добавление к источнику дополнительных гипертекстовых «смыслов» там, где они не подразумевались, иллюстраций, мультимедийного контента, какой-нибудь «Брежнев в комиксах» – это уже не добросовестное воспроизведение источника, а пародия. Смешение

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

опубликованных и неопубликованных (в то время) документов «в одном флаконе» тоже дезориентировало бы читателя.

Исходя из предположения, что оцифровка советского официоза в ближайшие годы так и останется уделом одиночек, не владеющих, чаще всего, языками программирования, можно обобщить типичные недостатки распространённых форматов хранения электронных версий этих изданий и выработать оптимальный формат, а скорее – несколько предпочтительных форматов для устранения этих недостатков или хотя бы минимизации наносимого ими ущерба.

2 Потребность в стандарте

Автору этих тезисов довелось принять участие в Круглом столе «Источники новейшего времени», проведённом Государственной публичной исторической библиотекой 27 сентября 2012 года. К тому моменту эксперименты с разметкой только начинались, стандарт существовал лишь в виде замысла – совокупности искомых параметров, изложенных ниже в разделе 4. Идея необходимости выработки стандартов представления оцифрованных книг и документов получила поддержку со стороны директора ГПИБ М. Д. Афанасьева и ряда других участников мероприятия.

2.1 Применимость стандарта

Речь идёт о стандарте не как о директивном документе, а как о наборе приёмов структурирования информации в удобном для использования и научного цитирования виде, которые создатель любого интернет-ресурса волен применять полностью или частично.

Вероятно, правильным будет выработать несколько стандартов Например, стандарт для параллельного воспроизведения текста на двух языках или для текста с многочисленными иноязычными вкраплениями будет отличаться от стандарта для моноязычных текстов; стандарт для текста, предусматривающего последующее редактирование (например, уточнение перевода или дополнение комментариев) – от стандарта для текстов, раз и навсегда «отлитых в граните». Необходим также стандарт для текстов, претерпевавших изменения в прошлом, или дошедших до нас в разных вариантах (аналог – правовая база «Консультант плюс» [9], позволяющая сравнивать все версии многократно изменявшихся правовых актов).

Мы имеем дело с советскими документальными источниками 1918–1991 годов, характерные особенности которых:

– современный русский алфавит;

– существование, как правило, единственного эталонного издания на бумаге, выпущенного массовым тиражом, и традиционной библиографической ссылки на издание, том, страницу;

– весьма ограниченный набор шрифтового и стилистического оформления (курсив, жирный шрифт, разрядка, подчёркивание), который легко воспроизводится средствами HTML;

– незначительное количество иллюстраций, диаграмм, формул, вкраплений иноязычных знаков (при этом собственно латинские буквы проблем не вызывают, некоторые сложности, впрочем – преодолимые, возникают с диакритикой).

2.2 Цели стандартизации

- взаимно однозначное соответствие между стандартной библиографической ссылкой на конкретную страницу либо конкретную главу издания (или конкретный текст в сборнике, томе, выпуске) и гипертекстовым адресом этой страницы или главы (этого текста);
- максимально простой и интуитивный формат адресации;
- возможность поиска текста в общеизвестной поисковой системе по любой цитате из него, решение проблемы искабельности цитаты, которая в оригинале «разорвана» служебной информацией (номером страницы, примечанием, колонтитулом, изображением), подходящее для распространённых браузеров;
- простая техника форматирования, которой можно обучиться за несколько часов, не будучи программистом по образованию.

3 Предшественники

Тема адекватного представления в Интернете источников прошлого обсуждается не первое десятилетие. См., например, стенограмму Круглого стола «Историк, источник и Интернет» (журнал «Новая и новейшая история», 2001, №2 [10]). Проблема «разрыва контекста» уже тогда затрагивалась в выступлении Т. Я. Валетова с кафедры исторической информатики МГУ, у которой есть собственная интернет-библиотека исторических источников [8].

Свой вариант формата – комбинированный текстово-графический разрабатывает С. И. Трифонов для Научно-педагогической электронной библиотеки РАО [17].

Стандарт для параллельного отображения текста на двух языках – Diglossa – разрабатывается Михаилом Быковым (Москва) [7]. Исходные тексты хранятся на сайте не в HTML, а в маркированном TXT. Версия же HTML генерируется в момент запроса с помощью JavaScript'a.

4 Параметры стандарта (техзадание)

4.1 Общие соображения о формате

Старейшие интернет-библиотеки (напр., Библиотека Максима Мошкова [1]) избегают сложных форматов, предпочитая Plain Text. Этот подход, однако, приводит к потере информации,

передаваемой несколькими нехитрыми полиграфическими приёмами: курсив, жирный шрифт, абзацные отступы, иерархия заголовков и т.п.

С нашей точки зрения, формат HTML [22] с минимальным добавлением CSS, с одной стороны, позволяет отразить особенности шрифтового оформления документальных источников, с другой – легко копируется в Word для дальнейшей работы с текстом без потери этого оформления (чего не скажешь о PDF [23] и DJVU [20]), с третьей – не требует никаких надстроек к интернет-браузеру. И, наконец, он идеально приспособлен для построения из текстов XX века системы взаимных ссылок.

4.2 Проблема адресации (URL)

Что сделает вебмастер-исполнитель, получив задание оцифровать в PlainHTML, к примеру, 9-томник Л. И. Брежнева «Ленинским курсом» [4]? Особенно если ему не пользоваться ресурсом, а лишь выполнить работу по его созданию.

В лучшем случае, назовёт файлы с томами как-нибудь вроде **lkurs4.html**. Создаст для них отдельную директорию. Может быть, даже проставит «якоря» типа **<A NAME="str9">** и ссылка «Л. И. Брежнев. Ленинским курсом, т.1, стр.9» будет выглядеть в лучшем случае так: **some-library.ru/brezhnev/lkurs/lkurs1.html#str9** . Тогда заказчик работы хотя бы сможет при необходимости исправить опечатку в файле с помощью простого текстового редактора.

Но более вероятно, что вебмастер воспользуется одним из языков программирования, соединив все тексты сайта в бинарную базу данных, работа с содержимым которой требует специальных навыков, а любое существенное изменение структуры подпадает под понятие «апгрейда» и требует длительных программистских усилий.

Наш общий замысел простого, интуитивного и лаконичного формата гиперссылки такой:

brezhnev.su/1/#02

(Л. И. Брежнев. 47-я годовщина Великой Октябрьской Социалистической революции. – в сб. Ленинским курсом, т.1 [текст №2 первого тома Сочинений]);

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

(Ведомости Верховного Совета СССР, 1991, № 52, ст. 1531);

pushkin.some-library.ru/4/#3

(А. С. Пушкин, Соч., т.4, с.3).

25съезд.кпсс.рф/1/#322

(XXV съезд КПСС. Стенографический отчет, т.1, с.322).

Традиционный подход к использованию человеко-понятного адреса ресурса (user-friendly URL или Clean URL) состоит в создании, для начала, адреса «человеко-непонятного» (типа **http://site.ru/index.php?book=lenin&vol=3&p=193**), формат которого задаётся языком программирования или системой контент-менеджмента, с последующим генерированием из него с помощью *Rewrite* или аналогичной команды, более внятного адреса для пользователя – человека [18]. Мы, в противовес этому, предлагаем сделать человеко-понятным собственно исходный URL (а не его псевдоним), непосредственно и исключительно средствами HTML.

Для этого каждая книга (том, выпуск) издания сохраняется в виде файла с именем **index.html** в отдельной директории, название которой состоит из одного числа – номера тома. Сразу после номера тома и знаков «слэш» и «решётка» указывается номер страницы. Номера страниц должны быть внедрены в текст в виде якорей типа **NAME** или **ID**.

Отметим, что в первом и втором примерах указаны реально существующие сайты (**brezhnev.su, sssr.su**), и что «интуитивность» доменного имени на практике ограничена доступностью на рынке или незанятостью конкретного домена второго уровня.

Заманчиво выглядит появившаяся в 2010 году возможность формировать полностью кириллические адреса (URLы). Однако, во-первых, это заметно усложняет работу для иностранных исследователей, которые могут читать кириллические буквы, но не всегда в состоянии ввести их с клавиатуры (та же проблема есть и в России – у пользователей русифицированных, вроде бы, мобильных устройств). Во-вторых, не каждый хостинг-провайдер полноценно работает с кириллическими названиями директорий. В-третьих, и браузеры довольно часто заменяют кириллические знаки кодами типа «%D0%9C%D0%B8…».

Выход видится в том, чтобы для каждого оцифруемого издания завести два домена – максимально «интуитивный» кириллический (сколь угодно высокого уровня, к примеру: 8съезд.народных.депутатов.рф) и параллельный латинский, а в «правой части» URLa (после первого слэша, следующего за доменным именем) избегать использования букв, как кириллических, так и латинских, ограничиваясь цифрами, знаками типа дефиса, присутствующего во всех версиях клавиатуры, и, по необходимости, знаком «решётки».

4.3 Проблема разорванной цитаты,

т.е. искабельности поисковыми системами последовательности слов, которая разорвана служебной информацией (номер страницы, колонтитул, рисунок, перенос слова) тоже имеет решение, хотя и не столь изящное.

Традиционные способы – указание номера страницы в квадратных скобках, указание его отдельной строкой – приводят к тому, что разорванная номером страницы цитата не ищется как единое целое ни Яндексом, ни Google.

5 Эксперименты с разметкой

Что касается собственно HTML как языка разметки, знакомство с его спецификацией [22] необходимо, но недостаточно для реализации стандарта на практике: В версии 5.1, которая ещё разрабатывается и уже внедряется, введено несколько новых тегов, не поддерживаемых пока ни одним браузером, в то же время из неё исключаются некоторые теги, которые давно и успешно применяются. Более того, даже в действующей с 1999 года версии 4.01 есть теги, которые ни один разработчик браузеров не счёл нужным задействовать (к примеру, font-stretch, который должен делать буквы более узкими или широкими – но не делает). А иные теги разными браузерами интерпретируются по-разному (например сочетание li и dd для одних пользователей создаёт абзацные отступы, а для других – превращает текст в «лесенку», сходящуюся к правому краю экрана).

Исходя из реалий сегодняшнего дня, а также, из надежд на здравый смысл производителей браузеров, которые вряд ли пойдут на поводу у разработчиков HTML в части исключения распространённых тегов и разрушат тем самым архитектуру значительного большинства ныне существующих сайтов, мы остановились на возможностях HTML 4.01, проверив конкретные варианты разметки на совместимость с пятью браузерами и двумя поисковыми системами.

Варианты разметки, позволяющей воспроизвести номера страниц и сноски, не разрушив цельность текста, тестировались на поисковых системах **Яндекс** [19] и **Google** [21], на браузерах **IE** 6, 7, 8 и 9, **Firefox** 3, 16 и 19, **Opera** 10, 11 и 12, **Chrome** 25 и 27, **Safari** for IPad.

Объектом эксперимента послужил первый том собрания сочинений Л. И. Брежнева – издание с довольно простой структурой: однородные заголовки трёх уровней; каждый текст занимает от одной до 30-40 страниц и начинается с новой страницы; почти отсутствуют символы расширенных кириллического и латинского алфавитов (изредка встречаются ударения и украинские буквы).

Вначале мы попытались воспользоваться тегом **<NOINDEX>...</NOINDEX>**, предложенным поисковой системой Яндекс для исключения фрагмента текста из индексации. Однако сочетание этого тега с другими, также необходимыми на границе страниц, сводило его эффект на нет: фрагмент до открывающего тега и после закрывающего не «склеивались» для поисковой машины в единый контекст. Кроме того, конкурирующая поисковая система Google такого тега не знает, что сокращает число российских пользователей, которые смогли бы найти размеченную им цитату, раза в полтора.

В дальнейшем испытывались в различных сочетаниях разметки, дающие следующий эффект:

- Номер страницы присутствует в виде всплывающей при наведении курсора на текст подсказки;
- Сноска присутствует дважды: сначала в виде такой же всплывающей подсказки и потом ещё раз в конце главы – в явном виде;
- Начало новой страницы, на какое бы место, включая середину перенесённого слова, оно ни пришлось, выделяется цветом текста или полурамкой вокруг первого слова (полуслова) страницы.

Испробовались:

а) 4 способа выделения первого (и, если необходимо, последнего) на странице слова (частей разорванного слова): **FONT COLOR; FONT STYLE; SPAN STYLE; U** (подчёркивание части слова).

б) 3 способа внедрения всплывающей подсказки с номером страницы: **A TITLE; SPAN TITLE; DIV TITLE**.

в) 10 способов проставления метки (якоря) номера страницы, в том числе:

– 3 варианта с тегом **A NAME ... /A** (тег открывается и закрывается в начале слова; открывается и закрывается в середине слова; открывается в середине слова, а закрывается в конце страницы);

– аналогичные 3 варианта с **A ID ... /A;**

– то же со **SPAN ID ... /SPAN**

– вставка атрибута **ID** в тег **FONT**, относящийся к первому слову новой страницы (или ко второй половинке разорванного слова).

После каждой очередной модификации разметки на экспериментальном сайте мы ждали, пока новая версия файла будет проиндексирована Яндексом и Google, затем проверяли на искабельность в обоих этих поисковиках все якоря и разорванные контексты.

Эксперимент с Яндексом постоянно выдавал трудноинтерпретируемые результаты: идентичные, казалось бы, контексты распознавались неодинаковым образом. Где-то поисковый робот вставлял в выдачу мнимую точку между словами или в середине слова, через которое проходил разрыв страницы – соответственно межстраничный контекст, заданный в кавычках в поисковой строке, Яндекс на странице не обнаруживал. В некоторых случаях помогало снятие кавычек (тогда он находил этот контекст как приблизительный), иногда и это не помогало (например, склеенное слово «обнаруживались» робот Яндекса не находил – для него на странице были слово «обна» и слово «руживались».

После каждой переиндексации следовал раунд переписки с программистской командой Яндекса, представители которой подчас сами не знали, на какое именно сочетание HTML-кодов поисковый робот реагирует тем или иным образом.

Сводные результаты эксперимента иллюстрируются следующей таблицей:

	Яндекс	**Google**	**IE 6, 7, 8, 9**	**Firefox 3, 16,19**	**Opera 10, 11, 12**	**Chrome 25, 27**	**Safari for IPad**	**Выводы**
1. Выделение границы страниц								
а) **FONT COLOR**	+	+	+	+	+	+	+	**OK**
б) **SPAN STYLE**	+	разрыв	+	+	+	+	+	–
в) **U / FONT**	+	+	+	+	+	+	+	**OK**
г) **FONT STYLE**	+	+	+	+	рамка слева не видна, если она в начале строки	+	+	**OK**
2. Метка страницы								
а) **<A NAME></A>** в середине слова	разрыв	+	+	+	+	+	+	–
б) **<A NAME TITLE>...</A>** в конце страницы	разрыв	+	+	+	+	+	+	–
в) **<A NAME></A>** в начале слова (паллиатив)	точка	+	+	+	+	+	+	–
г) **<A ID></A>** в середине слова	разрыв	+	+	+	+	+	+	–
д) **<A ID>...</A>** в конце страницы	разрыв	+	+	+	+	+	+	–
е) **<A ID></A>** в начале слова (паллиатив)	точка	+	+	+	+	+	+	–
ж) **<SPAN ID></SPAN>** в середине слова	разрыв	разрыв	+	+	+	+	+	–
з) **<SPAN ID TITLE>...</SPAN>** в конце страницы	разрыв	разрыв	+	+	+	+	+	–
и) **<SPAN ID></SPAN>** в начале слова (паллиатив)	точка	+	+	+	+	+	+	–
к) **FONT ID** в середине слова	+	+	+	+	+	+	+	**OK**
3. Всплывающий номер страницы								
а) **A TITLE**	разрыв (закрыва-ющим тегом **/A**)	+	+	+	10 только в 1-м абз. 11+ 12+	+	не виден	–
б) **DIV TITLE**	разрыв	разрыв	–	–	–	–	не виден	–
в) **SPAN TITLE**	+	разрыв	+	только в 1-м абз.	10+ 11, 12 только в 1-м абз.	только в 1-м абз.	не виден	–
г) **FONT TITLE**	+	+	+	+	+	+	не виден	**оптимален**

Итак, практика показала, что единственным «нейтральным» тегом, внедрить в который знак начала страницы и всплывающий её номер можно без ущерба для целостности цитаты, является тег **FONT**. Поскольку он же используется для графического выделения первого слова, тег приходится закрывать и тут же открывать повторно – теперь уже только с функцией **TITLE.**

Правда, настораживает тот факт, что из спецификации HTML 5 тег **FONT** предполагается исключить.

6 Текущая версия стандарта

На основании имеющегося опыта мы выбрали в качестве предпочтительного следующий вариант:

1. Выделяем уголком первое слово (полуслово) страницы, последнее слово (полуслово) предыдущей страницы выделять не будем.

2. В шапке файла создадим мини-таблицу стилей, в которой определим стиль **ps** (сокращение от PageStart), который будем указывать для первого слова (полуслова) каждой страницы. Для этого между тэгами **</TITLE>** и **</HEAD>** поместим следующий код:

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

(Пример 1)

(параметры стиля **ps**: рамка шириной в 1 точку; рамка сверху-есть, справа-нет, снизу-нет, слева-есть; цвет рамки – оранжевый).

Переход со страницы на страницу размечаем следующим образом:

```
марксизма-ленинизма, пролетар</font><font
id="17" title="17"
class="ps">ского</font><font title="17">
интернационализма,
```

(Пример 2)

Выглядит это так:

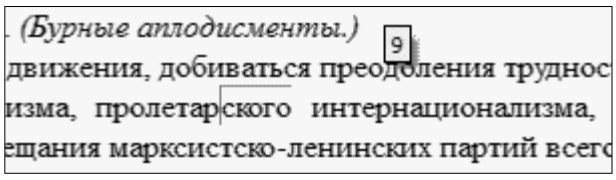

Рис.1

Если при распечатке нужно избежать воспроизведения на бумаге границ между страницами – это будет легко сделать, сохранив файл книги на локальном диске и вырезав из него текстовым редактором вышеприведённые строки таблицы стилей.

3. Некоторые нюансы для особых случаев:

а) Если страница начинается с таблицы, лучше создать отдельную пустую (состоящую из 10 жёстких пробелов) строку в левом верхнем углу новой страницы. Практика показывает, что получающийся уголок лучше видно, если эта строка имеет абзацный отступ и отделена от текста предыдущей страницы пустой строкой сверху.

б) Для страницы, начинающейся с заголовка, можно тоже создавать пустую маркированную строку или как обычно, выделять полурамкой первое слово, хоть оно и заголовок. Нам второй вариант нравится больше.

в) Если полуслово, с которого начинается страница очень короткое (1-2 буквы), можно распространить выделение полурамкой и на следующее слово. Но в этом случае лучше заменить пробел между ними жёстким пробелом: ** ** – чтобы выделенный рамкой контекст вдруг не разделился между двумя строками.

г) Если после выделяемого слова есть знак препинания, симпатичнее не включать его в выделение, остановившись (поставив замыкающий тег **</FONT>**) после последней буквы.

4. В книге, взятой нами для эксперимента, каждый текст (речь, статья) начинается с новой страницы, что позволяет нам в конце каждого текста вставить номер страницы явным образом:

```
<div
align="center"><br><i>10</i></div></font>
<a id="02"><hr></a>
<font id="11" title="11">
<h1>...
```

(Пример 3)

7 О степени трудоемкости исполнения

Вновь напомним, что призыв к стандартизации адресован прежде всего тем многочисленным энтузиастам, которые готовы к кропотливой ручной работе и уже занимаются ею, и лишь во вторую очередь – большим библиотекам, в которых объёмы оцифровки измеряются десятками и сотнями тысяч страниц.

При этом разметка правильно набранного (или распознанного) текста предложенным образом не кажется нам чересчур трудоёмкой. По сути, речь идёт о трёх-четырёх (не считая добавления единых для многотомника «шапки» и «хвоста» HTML-файла) блоках операций, которые можно в значительной степени автоматизировать использованием макрокоманд текстового редактора MS Word:

1) вставка между абзацами тегов конца и начала абзаца – ** ** ;

2) разметка заголовков трёх-четырёх уровней тегами **<h1>...</h1>** и т.п.;

3) замена номеров последней страницы каждого текста кодами, приведёнными выше в примере 3, а остальных номеров страниц – кодами из примера 2;

4) поиск и соответствующая html-разметка жирных, курсивных, латинских с диакритикой и иных нестандартных контекстов.

Вся работа по разметке отнимет несравнимо меньше времени, чем «ручная» вычитка (корректура) текста, без которой, по нашему мнению, обойтись нельзя. Ведь никакой спеллер не

увидит ничего удивительного в выражении, распознанном как ***«орава трудящихся»*** (вместо ***«права трудящихся»***) или ***«товарищ Терек»*** вместо ***«товарищ Герек»***. Иначе под видом распознанного текста мы получим такую же тарабарщину, какую имеем в русском разделе Google Books и других ресурсах, оцифрованных без применения человеческого интеллекта. При подобном количестве ошибок распознавания вопрос об адекватном поиске контекста ставить бессмысленно.

8 Заключение и перспективы

В данной версии доклада мы не воспроизводим внешний вид различных способов графического выделения границы страниц. Не приводим коды, выбранные для унификации стилей заголовков, не останавливаемся на частных случаях последовательности кодов для границы страниц, совпадающей с границей слов и абзацев. Всё это включено в полную версию описания стандарта на сайте **http://istnet.org/** , где будет также шаблон для создания типового файла книги и пошаговая инструкция для превращения исходного текста в HTML-файл, размеченный в соответствии со стандартом.

На очереди – выработка кроссбраузерного формата для оглавления (экспериментальный вариант уже применён в 1-м и 2-м томах Брежнева – **http://brezhnev.su/1/** и **http://brezhnev.su/2/**) и предметно-именных указателей.

Литература

[1] Библиотека Максима Мошкова. http://lib.ru/

[2] Библиотека «Нефть и газ». http://www.oglibrary.ru/

[3] Библия. Издание Московской Патриархии, М., 1988

[4] Брежнев Л. И. Ленинским курсом (в 9 т.). М., «Политиздат», 1970–1983.

[5] Военная литература. Электронная библиотека. http://militera.lib.ru/

[6] Геллер М., Некрич А. История России: 1917–1995. Утопия у власти. М., «МИК», 1995.

[7] Диглосса. Многоязычная библиотека. http://ru.diglossa.org/

[8] Исторические источники на русском языке в Интернете (Электронная библиотека Исторического факультета МГУ им. М. В. Ломоносова). http://www.hist.msu.ru/ER/Etext/

[9] КонсультантПлюс. Правовая информационная система. http://consultant.ru

[10] Круглый стол «Историк, источник и Интернет». // «Новая и новейшая история», 2001, №2. http://vivovoco.rsl.ru/VV/BONTONE/HISTORY.HTM

[11] Ленин В. И. Полн. собр. соч. Изд. 5-е (в 55 т.). М., «Политиздат», 1958–1965.

[12] Ленин В. И. Полное собрание сочинений. http://uaio.ru/vil/

[13] Маркс К., Энгельс Ф. Сочинения. Изд. 2-е (в 50 т.). Сайт Коммунистической партии Украины. http://www.kpu.ua/marx_engels_pss/

[14] Пихоя Р. Г. Советский Союз. История власти. 1945–1991. Новосибирск, «Сибирский хронограф», 2000.

[15] Правда, Петроград – Москва, 1917–1991.

[16] Солженицын А. И. Архипелаг ГУЛАГ. Опыт художественного исследования. Екатеринбург, «У-Фактория», 2006.

[17] Трифонов С. И. Комбинированное электронное представление печатных изданий. Сайт Всероссийской научной конференции RCDL-2011, Воронеж. http://rcdl.ru/doc/2011/paper51.pdf

[18] ЧПУ (Интернет) / Википедия http://ru.wikipedia.org/wiki/ЧПУ_(Интернет)

[19] Яндекс. Поисковая система. http://yandex.ru/

[20] DjVu 3. Спецификация. Неофициальный русский перевод. http://djvu-spec.narod.ru/djvu3-spec.pdf

[21] Google. Поисковая система. http://google.com

[22] HTML 4.01 Specification. W3C Recommendation 24 December 1999. http://www.w3.org/TR/html401/ Русский перевод (неофициальный): http://pyramidin.narod.ru/html401 HTML 5.1. A vocabulary and associated APIs for HTML and XHTML. W3C Working Draft 28 May 2013. http://www.w3.org/TR/html51/

[23] PDF Reference and Adobe Extensions to the PDF Specification http://www.adobe.com/devnet/pdf/pdf_reference.html

An Essay of Development of HTML-Based Standard of Electronic Publication for Digitizing of Soviet Source-Books

Grigory V. Belonuchkin

The article describes the results of the experiment with HTML code formatting to make the electronic copies of USSR officious books, such as “The Works” of CPSU leaders and legislatory bulletins, compatible with all the widely used Internet-browsers and popular search engines.

Other aims of elaboration of the unified HTML-coding Standard are: one-to-one correspondence of URL and bibliographic reference to the exact page of a book and its electronic version; the searchability of any quotation context, even if broken by service information (page number etc.).