

УДК 94(495), 004.9

ПРИМЕНЕНИЕ ТЕХНОЛОГИИ DOCSTOC ПРИ СОЗДАНИИ АННОТИРОВАННОЙ ЭЛЕКТРОННОЙ БИБЛИОТЕКИ О ВИЗАНТИЙСКОЙ ИМПЕРИИ

Шарапов Р.В.

*Муромский институт, филиал ФГБОУ ВПО «Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»,
Муром, e-mail: mivlgu@mail.ru*

В работе рассматриваются вопросы создания электронной библиотеки о Византийской империи. Обсуждается возможность использования технологии DocStoc для просмотра книг в режиме онлайн.

Ключевые слова: электронная библиотека, Византия, Византийская империя, DocStoc

USE THE DOCSTOC TECHNOLOGY IN ANNOTATED DIGITAL LIBRARY OF THE BYZANTINE EMPIRE

Sharapov R.V.

Murom Institute of Vladimir State University, Murom, e-mail: mivlgu@mail.ru

This paper considers the problem of creating a digital library of the Byzantine Empire. We discuss the possibility of use the technology DocStoc to view books online.

Keywords: digital library, Byzantine Empire, Byzantium, DocStoc

Проект «Византия» [1] был задуман как средство сведения разнообразной информации о Византийской империи (истории, культуре, архитектуре, искусству и религии) в одном месте. Важное место в проекте занимает электронная библиотека (<http://library.byzantium.ru/>) [2]. Она содержит книги, сборники и статьи о Византийской империи в PDF-формате. Каждая книга снабжена аннотированным описанием, включающим сведения об авторах и содержании книги. Пользователям предоставлена возможность просмотреть выходные данные и обложки книг [3].

Все книги в библиотеке представлены в виде графических копий в формате PDF. Это связано с тем, что большинство источников – это произведения XIX – начала XX века, полнотекстовых версий которых в современной редакции чаще всего не существует. Автоматическое распознавание подобных текстов (например, средствами Abbyy FineReader) также представляет собой проблему. Это связано с применением в этих произведениях, так называемой, дореволюционной орфографии. В результате наиболее приемлемым решением стало использование графических образов книг, полученных чаще всего путем сканирования оригиналов. Такие PDF-файлы занимают достаточно большое количество дискового пространства – размеры колеблются от 1 до 100 МБ.

Книги в электронной библиотеке снабжены описаниями, содержанием и аннотациями. Пользователь может свободно скачать каждую из книг на свой компьютер. Тем не менее при больших размерах книг,

пользователь может столкнуться с некоторыми неудобствами: узнать, подходит ли ему книга или нет, можно только скачав ее.

Возникает задача отображения книг на сайте до их полного скачивания. Такое отображение должно быть интерактивным, удобным для пользователя и экономичным (загружать только запрошенные пользователем страницы, а не всю книгу).

Первое решение, которое мы попробовали, было внедрение PDF-файлов на страницу. Но такое решение имело ряд недостатков. Во-первых, на компьютере пользователя должна быть установлена программа просмотра PDF (например, Adobe Reader). Во-вторых, при таком решении файл с книгой загружается целиком, что неудобно при больших его размерах.

Похожие решения можно реализовать с применением на сайте программ просмотра PDF-файлов, загружаемых на компьютер пользователя. Такие программы можно реализовать, например, с использованием Flash. Тем не менее подобное решение имеет все те же недостатки, что и раньше – файл с книгой загружается в них целиком (хотя, возможно и постепенно, по мере просмотра).

Ряд подходов позволяют сделать постраничный просмотр документов. Для этого каждая страница PDF-документа извлекается в отдельный графический файл (например, JPEG). Тогда, пользователь может быстро просмотреть нужные страницы, не тратя время на загрузку всего документа. Тем не менее при таком решении возникает необходимость дополнительного хранения каждой страницы книги в виде отдельного

файла. Размеры дискового пространства, занимаемого такими страницами, будут существенно превышать сами PDF-файлы.

Интересный подход предлагается проектом FlexPaper [4]. В данном случае документы конвертируются в SWF-файл (Flash-файл). Такой файл оптимизирован для просмотра в специальном FlexPaper-плеере. Решение достаточно удобно в использовании для небольших файлов. Тем не менее на сайте требуется хранить SWF-образы книг. В некоторых случаях такие файлы занимают неоправданно много места. Так, PDF-файл размером 50 Мб после преобразования в SWF-образы стал занимать более 350 Мб.

Достаточно интересное решение предлагает сервис Google. Документы [5]. Сервис позволяет хранить и просматривать постранично документы (хранящиеся как на сайте Google, так и на внешних ресурсах). Само по себе такое решение достаточно удобно. Тем не менее у сервиса было выявлено несколько существенных недостатков. Во-первых, он не позволяет просматривать PDF-файлы размером больше нескольких десятков мегабайт. Таким образом, большинство книг электронной библиотеки он не поддерживает. Во-вторых, документы с цветной палитрой (например, книги с желтым фоном) преобразуются в градации серого. Такое преобразование не всегда бывает оптимальным и иногда снижает читаемость информации.

Особый интерес представляет технология DocStoc [6]. Она позволяет загружать PDF документы на файловый хостинг DocStoc и просматривать их постранично через специальный интерфейс, встраиваемый на сайт. Имеется возможность работы и с другими форматами документов, например, с текстовыми файлами (DOC, DOCX, RTF, TXT), презентациям (PPT, PPS, PPTX, PPSX) и электронными таблицами (XLS, XLSX). Особенностью данной технологии является то, что новые страницы документа загружаются по мере необходимости. Поэтому для просмотра нескольких страниц нет необходимости загружать с сайта всю электронную книгу. Кроме того, книги хранятся на внешнем по отношению к сайту хостинге, что сокращает нагрузку на сервер.

Файловый хостинг DocStoc используется для хранения десятков миллионов документов и обеспечивает ежемесячно миллионы загрузок. Файлы могут быть об-

щедоступными или приватными (недоступными другим пользователям). Просмотр хранимых файлов может осуществляться как через сайт DocStoc, так и через интерфейс, встраиваемый на внешние сайты (рисунки). По желанию пользователя просматриваемый документ можно распечатать или скачать на свой компьютер. DocStoc имеет более 3 млн зарегистрированных пользователей.

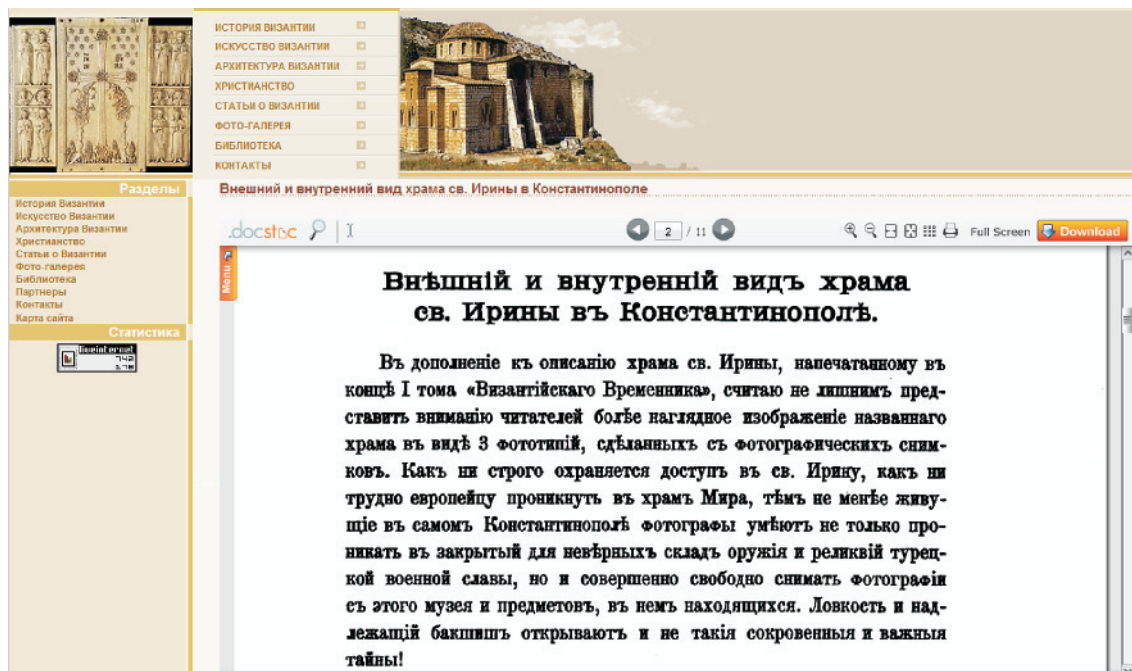
Несмотря на все преимущества, у DocStoc есть несколько особенностей, которые нужно учитывать при работе. Во-первых, документы масштабируются по ширине. Поэтому, если книга содержит вставки большего размера (фотографии, карты и т.д.), то за ширину документа будет принят наибольший размер таких вставок. Для устранения этого недостатка все книги должны быть нормированы по размеру. Во-вторых, цветные страницы, содержащиеся в книгах, автоматически преобразуются DocStoc в черно-белые, что приводит к потере их информативности.

Анализ показал, что применение технологии DocStoc является наиболее приемлемым вариантом при создании электронной библиотеки о Византийской империи (рисунок).

Таким образом, использование технологии DocStoc позволяет:

1. Обеспечить постраничный просмотр содержащихся в электронной библиотеке документов.
2. Существенно сократить потребление Интернет-трафика.
3. Избавиться от необходимости устанавливать дополнительное программное обеспечение (например, Adobe Reader) на компьютер пользователя (просмотр осуществляется через плагин DocStoc).
4. Сократить дисковое пространство, необходимое для размещения электронной библиотеки за счет хранения книг на файловом хостинге DocStoc.
5. Сократить издержки на реализацию проекта (использование DocStoc не требует никаких дополнительных затрат или лицензий).

В настоящее время аннотированная электронная библиотека доступна по адресу <http://library.byzantium.ru/>. Библиотека содержит несколько сотен книг и документов о Византийской империи. Для удобства пользователей предусмотрен поиск материалов по различным параметрам [7, 8].



Просмотр документов с использованием DocStoc

СПИСОК ЛИТЕРАТУРЫ

1. Византия – культура, история и искусство [Электронный ресурс]. – Режим доступа: <http://www.byzantium.ru> (дата обращения: 08.03.2012).
2. Шарапов Р. В. Свидетельство о регистрации электронного ресурса № 16883 от 24.03.11 «Электронный информационный ресурс Византия» // ОФЭРНИО.
3. Шарапов Р.В. Проблемы создания аннотированной электронной библиотеки о Византийской империи // Современные наукоемкие технологии – М.: Российская академия естествознания. – 2011. – № 6. – С. 67–69.
4. FlexPaper – The web based pdf viewer solution [Электронный ресурс]. – Режим доступа: <http://flexpaper.devaldi.com/> (дата обращения: 08.03.2012).
5. Документы Google [Электронный ресурс]. – Режим доступа: <http://docs.google.com> (дата обращения: 08.03.2012).
6. Docstoc – Documents, Templates, Forms, Ebooks, Papers & Presentations [Электронный ресурс]. – Режим доступа: <http://www.docstoc.com> (дата обращения: 08.03.2012).
7. Шарапов Р.В., Шарапова Е.В. Пути расширения булевой модели поиска // Информационные системы и технологии. Известия Орел ГТУ – Орел: ОрелГТУ, 2009. – №6(56) – С. 74–78.
8. Шарапов Р.В., Шарапова Е.В. Построение поисковых систем на основе теории нечётких множеств // Автоматизация и современные технологии. – 2011. – № 05. – С. 8–10.