XXXIII Неделя науки СПбГПУ. Материалы межвузовской научно-технической конференции. Ч.VI: С.29-30, 2005.

УДК 004.738.5

О.Д.Андреев (2 курс, каф. РТиТК), А.Б.Никитин, к.т.н., доц.

АВТОМАТИЗИРОВАННАЯ СИСТЕМА ПУБЛИКАЦИИ ГИПЕРТЕКСТОВЫХ НАУЧНО-ТЕХНИЧЕСКИХ ДОКУМЕНТОВ В WWW-СРЕДЕ

ABSTRACT: There are some problems related with technical documents publishing on the Internet. Ideology and techniques of publishing documents abounded with non-plain-text inline formulas, diagrams and illustrations were described. Powerful making-up application was developed.

При распространении в WWW-сети научных и учебных материалов технического характера авторы сталкиваются с проблемами публикации документов, насыщенных математическими формулами, графиками, чертежами. В основном, такие материалы подготовлены в doc-формате, в котором поддержка иерархии документов и перекрестных ссылок влечет за собой большой объем рутинной работы при обновлении документов.

В настоящее время существует множество универсальных систем, отличающихся различными подходами к публикации документов в Сети, однако среди них трудно найти удобную и подходящую для данного класса документов. Это связано с тем, что популярные системы не поддерживают редактирование текста и формул даже на уровне MS Word + MS Equation, а специализированные математические, схемотехнические и прочие системы не предназначены для полноценной поддержки документов в Сети. (В качестве типичной системы первого типа можно привести пример CityDesk [1], в качестве второй – Mathematica [2]). В связи с этим, имеется необходимость в разработке и исследовании новой системы публикации, учитывающей особенности научно-технических документов.

В ходе создания обозначенной системы были поставлены и решены следующие задачи:

- описание принципов обработки внутристрочных (inline) формул в виду их переноса в HTML-формат;
- разработка синтаксиса проставления гиперссылок на документы, формулы и иллюстрации, а также синтаксиса разметки иерархии рубрик и ключевых слов для алфавитного указателя;
- разработка программы, генерирующей готовые HTML-страницы;
- создание документации по синтаксису разметки и по работе с программой.

Разработанное ПО относится к разряду клиентских систем управления содержимым сайта, которые хранят исходные данные на стороне клиента и отправляют на сервер только готовые HTML-страницы. Таким образом, от сервера не требуется специфическое ПО, такое, как PHP или ASP для того, чтобы производить операции по «сборке» страниц из исходных данных.

Требования, предъявляемые к автоматизированной системе состояли в следующем.

1. Результат публикации есть веб-сайт, основанный на статических HTML-файлах.
2. Документы имеют иерархическую структуру со стандартной для научных изданий нумерацией вида «5.1.3».
3. Каждый раздел сайта должен быть доступен для скачивания в виде ZIP-архива.
4. Сайт должен иметь алфавитный указатель в привычном для печатной литературы виде.
5. Максимальная простота обновления иерархической структуры и содержимого документов.

На сегодняшний день не существует распространённого и удобного формата описания формул в рамках (X)HTML [3,4], поэтому было решено хранить их в виде GIF-картинок, получающихся после экспорта из формата Word. Логическая структура при этом не страдает,

т.к. сайт представляет собой лишь статичный результат публикации, а структура содержится в исходных Word-файлах.

В качестве технологической базы для разработки системы использовался язык PHP [5]. Система включает в себя PHP-сценарий (запускается из веб-браузера), набор документов-полуфабрикатов и файл Tree.txt, в котором, в текстовом формате хранится дерево документов, а также наборы ключевых слов для создания алфавитного указателя.

Алгоритм процедуры публикации состоит в следующем.

1. Редактируется индексный файл Tree.txt, в котором описывается иерархия документов. Если у документа есть дочерние документы, то он условно считается разделом и на его HTML-странице есть дерево дочерних документов с возможностью скачать offline-версию. В индексном файле описываются заголовок, идентификатор для гиперссылок и набор ключевых слов для каждого документа.
2. Массив Word-документов, сверстанных в соответствии с разработанными правилами разметки, экспортируется с помощью макроса VisualBasic в массив HTML-файлов, насыщенных излишними метатегами. Имя и местоположение каждого документа определяется в индексном файле Tree.txt. За основу разметки исходного Word-документа взят т.н. Wacko-синтаксис [6] – система простых правил текстовой разметки.
3. Запуск PHP-сценария в режиме создания «полуфабрикатов» – очищение HTML-файлов, удаление из них любого оформления, чтобы остались лишь тело документа и спецразметка.
4. Второй режим работы сценария собирает из полученных полуфабрикатов полноценные статические страницы, обрабатывает разметку, превращая её в HTML-теги, автоматически расставляет вложенную нумерацию.
5. Публикация HTML-файлов на сервере с помощью FTP или иных протоколов.

Разработанная автоматизированная система показала себя удобной в практическом использовании и вполне понятной для неподготовленного пользователя. Опробованные при создании системы архитектурные решения могут быть использованы в более сложной и универсальной системе публикации, работа над которой ведётся в настоящее время.

ЛИТЕРАТУРА:

1. CityDesk, http://fogcreek.com/CityDesk/index.html
2. Mathematica, http://wolfram.com
3. Presenting mathematical expressions on Web pages, Jukka Korpela, http://cs.tut.fi/~jkorpela/math/
4. The State of MathML: Mathematically Speaking (and Stuttering), Pankaj Kamthan, http://tech.irt.org/articles/js208/index.htm
5. Официальный сайт PHP, http://php.net/
6. Идеология и синтаксис Wacko-разметки, http://www.npj.ru/npjdev/articles/razmetka