Архитектура и реализация системы управления контентом Internet-библиотеки CitCMS

© Е.Е. Сальникова[1], С.А. Сальников[1], С.Д. Кузнецов[2]

[1]ЦИТФорум, [2]ИСП РАН, г. Москва

elev@citforum.ru, serg@citforum.ru, kuzloc@ispras.ru

Аннотация

На конференции RCDL'2009 был представлен доклад [1], в котором обосновывалась потребность к созданию новой системы управления контентом для поддержки полнотекстовых научно-технических Internet-библиотек (Web Content Management System, WCMS), формулировались требования к такой системе и обсуждались технологии, которые можно использовать при ее построении. За прошедший год было проведено несколько экспериментов, построено несколько опытных вариантов WCMS, которые испытывались на реальном контенте. В данной статье описываются выводы, которые можно сделать на основе полученного опыта, и обсуждаются основные компоненты, используемые в результирующей WCMS CitCMS.

1 Введение

Чем дальше, тем больше веб становится основным источником информации в различных областях науки, техники, производственной деятельности, искусства и т. д. С каждым годом в Internet доступно все большее число публикаций журнальных статей, книг, а также материалов, которые написаны специально для электронной публикации (например, блогов). Будучи одним из продуктов информационной технологии, Web активно способствует ее дальнейшему развитию, IT-публикации все в большей степени перекочевывают в среду Internet. Хорошими примерами являются электронные библиотеки ведущих мировых компьютерных сообществ IEEE Computer Society [2] и Association for Computer Machinery (ACM) [3], в которых содержатся материалы всех журналов, издаваемых этими сообществами, и труды ведущих мировых конференций. Во многих случаях труды этих конференций вообще не издаются в бумажной форме, доступны только в Web и вовсе не утрачивают при этом авторитетности.

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

Одним из показательных примеров Российского сегмента мирового компьютерного интернет-сообщества является Internet-библиотека информационных технологий CITForum [4]. Эта библиотека существует уже более 15 лет и содержит публикации разного объема (от небольших заметок до крупных книг), посвященные различным аспектам информационных технологий: теория и средства программирования, операционные системы, сетевые технологии, системы баз данных, информационные системы и т. д. Контент библиотеки непрерывно развивается и обновляется за счет как перепечатки (законным образом) материалов других компьютерных изданий, так и публикации оригинальных материалов, специально написанных для CITForum. Оба вида публикаций полностью себя оправдывают. В частности, накопление в библиотеке «вторичных» материалов позволяет обеспечить их доступность и надежность хранения, а публикация «первичных» статей и книг позволяет предельно быстро донести их до читательской аудитории.

Аудитория CITForum широка и разнообразна: преподаватели университетов и вузов, аспиранты и студенты; начинающие программисты, системные администраторы и менеджеры проектов; профессионалы в области информационных технологий и т. д. Число подписчиков, которым регулярно рассылаются новости библиотеки, составляет около 50 тысяч. На публикации библиотеки CITForum имеются многочисленные ссылки в других журнальных статьях и книгах.

Как и во многих других интернет-изданиях, образовавшихся на заре веб-технологий, в библиотеке CITForum материалы публикуются в формате HTML, что делало и делает их легко доступными на любых веб-клиентах, позволяет использовать любые операционные системы и браузеры, не требует большой пропускной способности сети и т. д. Однако эта полная изначальная ориентация на HTML привела к тому, что публикуемые материалы и сохраняются в формате HTML. По мере развития библиотеки, появления многочисленных форматов, в которых представляются перепечатываемые материалы (XML с таблицами стилей, TeX, PDF, Word и т. д.), повышения уровня требований к редакторскому качеству публикаций и т. д. «унаследованная» от эпохи становления веба примитивная система управления контентом становится все более обременительной для редакторов, администраторов

и авторов публикаций. Следует отметить, что эта ситуация свойственна всем долгоживущим Web-репозиториям публикаций, из которых в Рунет в области информационных технологий, кроме CIT-Forum, можно выделить, например, электронные библиотеки издательства «Открытые системы» [5] и компании «Интерфейс» [6], создателям которых также приходилось сталкиваться с подобными проблемами. Как отмечалось в [1], по мере развития Internet-библиотеки:

- все труднее обеспечивать ее должное качество;
- велика трудоемкость подготовки первичных и вторичных материалов к публикации;
- трудно менять рубрикацию материалов, вводить новые разделы и т. д.;
- нелегко обеспечивать абсолютную гарантию сохранности и доступности ранее опубликованных материалов;
- все более сложной становится задача общего администрирования библиотеки.

В нашем докладе [1] на конференции RCDL'2009 приводился и обосновывался полный список требований к WCMS, пригодной для качественной поддержки современной Internet-библиотеки, пояснялись причины, по которым для этих целей невозможно использовать существующие коммерческие и свободно доступные системы управления контентом, а также обосновывался выбор технологий, на основе которых можно было бы построить такую WCMS.

За прошедший год при выполнении проекта♣ было проведено несколько экспериментов, построено несколько опытных вариантов WCMS, которые испытывались на реальном контенте библиотеки CITForum. Были выявлены наиболее настоятельные потребности, и для их удовлетворения была построена первая действующая версия WSMC CitCMS [7], которая находится в состоянии опытной эксплуатации. Общая архитектура системы позволяет безболезненно ее расширять, добиваясь удовлетворения остальных требований, сформулированных в [1].

Основная часть статьи организована следующим образом. В разделе 2 обсуждается сложившийся сценарий поддержки библиотеки CITForum и выявляются основные возникающие проблемы. В третьем разделе описывается общая архитектура разработанной WCMS CitCMS и обосновывается использование технологий, на которых она базируется. Раздел 4 содержит заключение, в котором рассматривается текущее состояние дел и описываются планы на ближайшее будущее.

2 Сложившийся сценарий поддержки библиотеки CITForum и основные проблемы

Основную массу хранимых и публикуемых документов библиотеки CITForum составляют технические и аналитические статьи и книги. Эти материалы присылаются авторами («первичные» публикации) или отбираются редакцией библиотеки среди материалов, уже опубликованных в электронном виде другими Internet-изданиями («вторичные» публикации). В последнем случае, естественно, соблюдаются все формальности, связанные с защитой интеллектуальной собственности и авторских прав.

Поскольку CITForum – это библиотека информационных технологий, и публикуемые материалы носят технический характер, публикации обычно имеют достаточно большой объем и сравнительно четкую структуру. В основном документы поступают в редакцию в формате HTML или каком-либо варианте Word. Их обработка сводится к примерно одинаковому форматированию текстов с применением ряда существующих скриптов, после чего выполняется небольшое «причесывание» получаемого HTML-представления. Основная проблема этой части сценария состоит в том, что в ряде случаев авторы материалов позже присылают исправленные версии своих материалов, и в этом случае отсутствует возможность автоматически учесть эти исправления в имеющихся «причесанных» версиях материалов.

Зачастую материалы готовятся к публикации совместно авторами и редакторами CITForum. Для этого в настоящее время приходится использовать «ручной» механизм синхронизации доступа к документам на сервере, что неудобно и авторам, и редакторам, а также часто приводит к возникновению трудно исправляемых ошибок. Особенно затруднена совместная работа над материалами большого объема, такими, как книги и руководства. Для устранения этой проблемы требуется механизм, поддерживающий удобную и надежную коллективную работу над документами, вообще говоря, произвольно большого объема.

Решению этих первоочередных проблем и посвящается первая версия разрабатываемой WCMS CitCMS.

3 Архитектура и основные компоненты системы

В общей архитектуре разработанной версии WCMS CitCMS обеспечиваются компоненты, поддерживающие хранение документов, их совместную подготовку, импорт документов, подготовленных с использованием других средств, форматирование документов для их публикации и редактирование крупных документов.

3.1 Основа системы – ikiwiki

В [1] отмечалось, что следует более детально исследовать возможность использования в качестве основы требуемой WCMS какую-либо из систем поддержки Wiki. В результате изучения и сравнения разных подобных систем, а также создания действующих прототипов WCMS выбор пал на систему ikiwiki [8], разрабатываемую с 2006 г. сообществом open source под руководством Джоуи Хесса (Joey Hess), одного из ведущих участников проекта Debian GNU/Linux [9].

Почему мы решили использовать wiki вообще? Ответ на этот вопрос прост: основная функция, поддерживаемая wiki, а именно, функция поддержки коллективного редактирования документов в режиме онлайн с сохранением истории изменений, очень близка к основной функции, требуемой от WCMS. Другими словами, при создании новой WCMS целесообразно как можно более основательно воспользоваться возможностями существующих wiki.

Почему мы выбрали именно ikiwiki? На этот вопрос можно дать несколько ответов. Во-первых, ikiwiki отличается стилем своей разработки (так называемый подход Unix – Unix Way). У ikiwiki имеется собственное ядро небольшого размера, и широко используются другие существующие свободно распространяемые инструменты. Допускается простое и естественное расширение функциональных возможностей системы.

Как уже отмечалось, система является **расширяемой** в силу самой своей организации. Практически все функциональные возможности последних версий системы реализуются в виде подключаемых модулей (плагинов). В частности, на основе реализации соответствующих плагинов поддерживается несколько входных форматов документов. При потребности можно достаточно просто реализовать новые плагины для поддержки дополнительных входных форматов.

Во-вторых, доверие вызывает команда разработчиков ikiwiki, руководитель которой, Джоуи Хесс, хорошо зарекомендовал себя в проекте Debian. В мире open source надежность команды часто играет не меньшую (а иногда и большую) роль, чем качество самого кода. К сожалению, любой программный продукт с открытыми кодами почти неминуемо обречен на исчезновение, если его перестают сопровождать и/или развивать исходные разработчики.

Наконец, у ikiwiki имеется ряд технических особенностей, делающих эту систему особенно привлекательной именно в нашем случае. В отличие от почти всех остальных wiki и WCMS вообще, для хранения документов и их версий используются зрелые и полнофункциональные **системы управления версиями** (Version Control System, VCS). В ранних версиях системы допускалось использование только VCS Subversion [10], но впоследствии за счет достаточно простого и общего интерфейса с VCS и внедрения механизма плагинов появилась возможность использования разных VCS. В частности, в настоящее время среди пользователей ikiwiki большой популярностью пользуется VCS git [11]. Впрочем, как будет показано в следующем разделе, для нашего проекта более предпочтительной является VCS Subversion.

Важной особенностью ikiwiki является то, что результирующие HTML-страницы образуются путем **компиляции** внутреннего представления контента, а не генерируются динамически. При работе с достаточно крупными документами для генерации соответствующей HTML-страницы может потребоваться несколько секунд, что в большинстве случаев неприемлемо для пользователей Web-сайтов. В других WCMS с этой проблемой борются с использованием кэширования данных, но при использовании техники компиляции такая проблема просто не возникает.

Еще одной отличительной особенностью ikiwiki является то, что вся система и основные плагины написаны на **языке Perl**. Как отмечалось в [1], имеется ряд доводов в пользу использования Perl при разработке WCMS. В данном случае наличие огромного репозитория CPAN [12] различных программ, написанных на языке Perl, существенно упрощает создание дополнительных плагинов, которые требуются для получения целевой WCMS.

3.2 Хранение документов

Во многих WCMS, предназначенных для управления коллекциями небольших документов, для хранения контента и поддержки его версий используются SQL-ориентированные СУБД (при этом результирующие HTML-страницы генерируются динамически). Однако в Internet-библиотеках, подобных CITForum, зачастую сохраняющих документы объемом в несколько мегабайт, над которыми выполняются сотни незначительных правок, применение СУБД оказывается неэффективным. Для этого в большой степени подходят системы управления версиями.

Кажется естественным использование одной из проверенных временем, «зрелых» VCS, таких, как Subversion [10], git [11], Mercurial [13], Bazaar [14]. Все эти системы обеспечивают с нашей точки зрения примерно одинаковые функциональные возможности и различаются в деталях.

В настоящее время совершенно неочевидно, что в нашей WCMS могут потребоваться какие-либо особые возможности распределенных VCS, к которым относятся VCS git, рекомендуемая сейчас в документации ikiwiki как «вариант по умолчанию», Mercurial и Bazaar. И поэтому в первой версии CitCMS используется VCS Subversion. Эта система является самой зрелой и одной из самых распространенных из всех VCS с открытыми исходными кодами, и её возможностей вполне хватает для удовлетворения потребностей WCMS.

Впрочем, как отмечалось выше, в ikiwiki, на которой основана CitCMS, можно использовать разные системы управления версиями, так что ничто не мешает в случае надобности поменять Subversion на другую VCS в следующих версиях CitCMS.

3.3 Импорт документов

Требуемые для CitCMS возможности обеспечиваются за счет создания новых плагинов для ikiwiki. В текущей версии CitCMS поддерживается два формата входных документов, наиболее актуальных для библиотеки CITForum – Word и HTML.

Word – это очень распространенный текстовый редактор, многие публикации поступают именно в

этом формате, и основная проблема состоит в том, что в общем случае тексты, представленные в формате Word, просто невозможно автоматически переводить в формат структурированного текста в смысле, принятом в сообществе WWW: используются разные модели представления документов. Проблема усугубляется сложностью и закрытостью формата Word и наличием связанных с этим ошибок в свободных (и проприетарных!) реализациях. В ряде случаев удается преобразовать документы в формате Word к формату CitCMS за счет использования средств OpenOffice.org [15] и собственных программных средств, но в общем случае без доли ручного труда обойтись не удается.

В очень многих случаях публикация, которую требуется отредактировать и разметить в библиотеке CITForum, изначально представлена в формате HTML. В этом случае CitCMS пытается автоматически привести исходный формат HTML к своему внутреннему формату. Общего алгоритма преобразования не существует, но поддерживаются наиболее распространенные варианты исходного HTML-представления (при отклонении от них опять же требуется вмешательство человека).

Частным, но важным случаем задачи импорта документов, представленных в формате HTML, является перенос в среду CitCMS текущего контента библиотеки CITForum. Это объемная и непростая работа, поскольку контент накапливался в течение долгого времени без точного отслеживания стандартов HTML-представления документов, но она необходима для CITForum и чрезвычайно полезна для тестирования CitCMS.

3.4 Форматирование

Поскольку система ikiwiki в своем исходном виде применяется для публикации документов в стиле Wiki, в CitCMS требуется добавление средств форматирования, естественных для Internet-библиотеки. Такие средства форматирования реализуются в виде набора дополнительных плагинов и поддерживают, в частности, возможности разбиения крупных документов на страницы, автоматическое оформление ссылок на используемые литературные источники, обработку крупных изображений и т. д.

4 Заключение

В первой работоспособной версии новой WCMS CitCMS решен ряд важных задач:

- спроектирована общая архитектура системы;
- выбраны основные готовые компоненты, которые можно использовать в реализации;
- проведена первая фаза опытной эксплуатации.

Система является естественным образом расширяемой. Хотя в настоящее время она позволяет решать только наиболее важные задачи управления контентом научно-технической Internet-библиотеки, в дальнейшем возможно наращивание ее возможностей для достижения всех целей, сформулированных в [1]. Кроме того, система опирается на использование хорошо поддерживаемых готовых программных средств с открытыми исходными кодами, и сама будет распространяться аналогичным образом после достижения требуемого уровня зрелости. Это позволит использовать ее в ряде других проектов, требующих поддержки веб-контента.

Литература

[1] Кузнецов С.Д., Сальникова Е.Е., Сальников С.А. Управление контентом в крупных научно-технических Internet-библиотеках// Труды XI Всерос. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Петрозаводск, 17 – 21 сентября 2009 г.

[2] IEEE Computer Society Digital Library. – http://www.computer.org/portal/web/csdl, 2010.

[3] The ACM Digital Library. – http://portal.acm.org/dl.cfm, 2010.

[4] Библиотека CITForum. – http://citforum.ru/, 2010.

[5] Wiki проекта CitCMS. – http://citcms.org/, 2010.

[6] Web-сайт издательства «Открытые системы». – http://www.osp.ru/, 2010.

[7] Веб-сайт компании «Интерфейс». – http://www.interface.ru/, 2010.

[8] Wiki проекта ikiwiki. – http://ikiwiki.info/, 2010.

[9] Веб-сайт проекта Debian. – http://www.debian.org/, 2010.

[10] Веб-сайт проекта Apache Subversion. – http://subversion.apache.org/, 2010.

[11] Веб-сайт проекта git. – http://git-scm.com/, 2010.

[12] CPAN: Comprehensive Perl Archive Network. – http://www.cpan.org/, 2010

[13] Веб-сайт проекта Mercurial. – http://mercurial.selenic.com/, 2010

[14] Веб-сайт проекта Bazaar. – http://bazaar. canonical.com/en/, 2010.

[15] Веб-сайт проекта OpenOffice.org. – http://www.openoffice.org/, 2010.

Architecture and implementation of the Content Management System CitCMS for internet library

E.E. Salnikova, S.A. Salnikov, S.D. Kuznetsov

The paper [1] presented at the RCDL'2009 motivated needs for a new content management system to support full-text scientific and technological Internet libraries (Web Content Management System, WCMS), stated requirements to such a system, and discussed technologies that might be used in it's implementation. During last year several experiments has been conducted, several prototype systems have been implemented and tested based on real content. This paper describes results and conclusions of our experiments, and discusses main components used within a resulted WCMS CitCMS.

♣ Работа выполнена при финансовой поддержке РФФИ (проект 09-07-00282)