ЭВОЛЮЦИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ СИСТЕМЫ ПОДГОТОВКИ МАТЕРИАЛОВ ДЛЯ ЭЛЕКТРОННОЙ БИБЛИОТЕКИ «НАУЧНОЕ НАСЛЕДИЕ РОССИИ»

Погорелко К.П.
(Библиотека Математического института им. В.А.Стеклова РАН — отдел БЕН РАН)

Программное обеспечение электронной библиотеки «Научное наследие России» реализовано в виде нескольких независимых систем, которые, взаимодействуя друг с другом, обеспечивают различные технологические процессы функционирования проекта. В данной работе рассматриваются вопросы эволюции системы, обеспечивающей подготовку электронных публикаций. Эта система обеспечивает участникам проекта ввод отсканированных изображений оригинального документа, возможность формирования системы навигации в виде иерархического оглавления и дает возможности выпускающей группе контролировать ход процесса и качество представляемого материала. Для публикации электронных документов в Интернет готовые документы экспортируются в систему обслуживания читателей, которая позволяет пользователям осуществлять поиск и просматривать найденные публикации.

Первый вариант программного обеспечение для подготовки электронных публикаций создавался в 2007 г. на базе технологий электронной библиотеки Математического института им. В.А. Стеклова РАН и соответствовал требованиям, предъявляемым к системе на тот момент времени [1-3]. К настоящему времени произошли изменения, вызванные как совершенствованием технических средств, участвующих в процессе подготовки электронных публикаций, так и уточнениями и изменениями технологического процесса.

Основные временные и материальные затраты в процессе подготовки электронных публикаций приходятся на процесс оцифровки первоисточников. Кроме того, процедура оцифровки, так или иначе, влияет на печат-

ный оригинал. Поэтому необходимо обеспечить получение на имеющихся технических средствах максимально возможного качественного результата, с учетом необходимой производительности труда, чтобы в будущем не пришлось бы возвращаться к повторному сканированию первоисточников. Первоначально документы сканировались, в основном, в черно-белом формате, а изображения в градациях серого или цветные являлись большей частью исключением и формировались на планшетных сканерах, которые не обеспечивали необходимой производительности. Кроме того, были ограничения и на объем памяти, имевшийся для хранения архивов. В настоящее время в проекте задействованы книжные сканеры, позволяющие оцифровывать первоисточники в цветном формате. Также в последнее время значительно увеличены объемы дискового пространства для хранения электронной библиотеки. Эти изменения ставят на повестку дня необходимость модификации программного комплекса системы подготовки электронных публикаций для обеспечения работы с цветными копиями первоисточников.

Увеличение объемов обрабатываемой информации потребует, прежде всего, перехода на более мощный сервер, что может вызвать определенные проблемы при переносе существующего программного обеспечения. Это связано с тем, что модули, используемые в системе для просмотра загруженных изображений и получения PDF файлов, в целях эффективности были реализованы на языке C++ для платформы x-32. Поэтому для перехода на более мощную платформу потребуется переделка этих модулей на платформу x-64 или процессорно-независимую платформу .NET.

Передача информации от участников проекта в центральное хранилище происходит в системе по протоколу HTTP, который, в свою очередь, использует протокол TCP. Однако, когда объем файлов значителен, а линии связи ненадежны, средств коррекции трафика, реализованных в протоколе TCP, оказывается недостаточно. В настоящее время это приводит к обрыву соединения TCP, зависанию процедуры обмена и к необходимости начинать загрузку файлов с изображениями заново. Переход

на цветной формат приведет, прежде всего, к увеличению объема передаваемой в систему информации, что потребует изменения программного обеспечения, обеспечивающего передачу файлов. Необходимо добавить возможность, при которой большие файлы могли бы передаваться по частям, и, в случае обрыва связи, продолжать загрузку с прерванного места.

Следующим направлением изменений программного обеспечения является более гибкое обеспечение работы группы выпуска документов. В первоначальном варианте технологической цепочки по подготовке электронных публикаций предполагалось, что участники проекта загружают на сервер уже готовые электронные публикации, в которых качество отсканированного материала соответствует требованиям проекта. На выпускающую группу была возложена только функция контроля. Однако в ходе развития проекта появились участники, которые не в состоянии сами обеспечить необходимое качество отсканированного материала. Сложилась практика, при которой выпускающая группа стала проводить обработку загруженных изображений с целью их улучшения. В настоящее время реализована возможность исправления одиночных файлов. Для исправления всей публикации используется возможность загрузки с сервера подготовки изображений на рабочие компьютеры выпускающей группы электронной публикации в формате PDF и повторной загрузки на сервер исправленного набора файлов. Это не совсем удобно при существующих объемах информации и станет определенной проблемой при увеличении объемов, связанных с переходом к цветным форматам. Поэтому возникает необходимость в улучшении реализации программного обеспечения для выпускающей группы, которое позволит заменять произвольное количество файлов электронного документа.

В настоящее время единственным форматом графических файлов, с которым работает система, является формат TIFF. В этом формате хорошо обеспечивается работа с черно-белыми файлами, однако для хранения файлов в градациях серого или цветных этот формат не является лучшим, так как форматы компрессии этих файлов не закреплены стандартом. Поэтому возникает

необходимость обеспечения в системе возможности комплектования электронной публикации из файлов разных форматов. Это потребует определенных изменений как в структуре базы данных электронных публикаций, так и в программном обеспечении, поддерживающем работу с изображениями. Однако основной проблемой для такого перехода будет изменение взаимодействия с системой обслуживания читателей, которая так же, как и существующий вариант системы подготовки электронных публикаций, рассчитана на работу с файлами одного формата. Предполагается при экспорте готовых документов в систему обслуживания читателей производить графическое преобразование загруженных изображений в единый графический формат PNG и приводить их к единой плотности 200 точек на дюйм. Однако, наилучшим решением на взгляд автора являлось бы решение, при котором система просмотра электронных публикаций была бы выведена из системы обслуживания читателей в качестве самостоятельной системы. Такое решение позволило бы развивать систему просмотра электронных публикаций независимо от остальной системы обслуживания читателей и обеспечить качество предоставляемых услуг в соответствии с современными требованиями пользователей.

Литература

1. *Погорелко К.П. Вопросы создания полнотекстовой базы данных в Библиотеке Математического института им. В.А. Стеклова РАН // Информационное обеспечение науки: новые технологии: Сб. науч. тр. под ред. Н.Е. Каленова — М.: БЕН РАН, — 2005. — С. 270-274.*
2. *Погорелко К.П. Комплекс программ для создания полнотекстовой электронной библиотеки // Новые технологии в информационном обеспечении науки: : Сб. науч. тр. под ред. Н.Е. Каленова — М.: Научный мир, — 2007. — С. 66-68.*
3. *Нестеренко А.К., Сысоев Т.М., Погорелко К.П. Задача реализации электронной библиотеки "Научное наследие России" как распределенной информационной системы // Новые технологии в информационном обеспечении науки: : Сб. науч. тр. под ред. Н.Е. Каленова — М.: Научный мир, — 2007. — С. 276-287.*