Технологический процесс подготовки изданий на примере Фундаментальной электронной библиотеки «Русская литература и фольклор». Текущее состояние и принципы модернизации

© С. И. Трифонов, А.Е. Поляков

ФУГП НТЦ «ИНФОРМРЕГИСТР»
trf@ya.ru, pollex@mail.ru

Аннотация

В докладе описывается технология перевода изданий из печатного вида в электронную форму, разработанная и внедрённая в отделе электронных библиотек ФУГП НТЦ «ИНФОРМРЕГИСТР», в частности, в Фундаментальной электронной библиотеке «Русская литература и фольклор» (http://www.feb-web.ru)
Обсуждаются как текущее состояние технологии, так и принципы проводящейся в настоящее время модернизации
Перевод изданий из печатного вида в электронную форму представляет собой трудоёмкую задачу, крайне актуальную для библиотечного сообщества. Далеко не все стадии подготовки электронного массива поддаются полной автоматизации, поэтому качественно подготовленный документ оказывается во многом результатом ручного труда. Цель наших разработок - создание программного окружения, оптимизирующего ручные процедуры, в котором трудоёмкий полуавтоматический процесс подготовки документов выполнялся бы максимально удобно, просто и быстро.

1 Структура электронного издания

Существуют по крайней мере несколько способов оформления документа в электронном виде. Технология, обсуждаемая в докладе, предназначена для подготовки документов в формате HTML, с полноценной отработкой структурных элементов, стандартных для традиционных печатных изданий: иллюстрации, сноски, разбиение на страницы. Результирующий электронный документ включает также важную информацию, содержащуюся в исходном документе условно, но не достаточно формально. Это - логическая структура документа, отображающаяся обычно в виде оглавления, а также ссылки между документами и фрагментами.

Все электронные издания и произведения, которые готовятся по технологии, в обязательном порядке снабжаются библиографическими описаниями в формате, соответствующем стандарту ГОСТ 7.1-2003.

2 Процесс подготовки электронных изданий

По технологическим причинам, процесс производится раздельно для каждого издания. Как правило, под изданием подразумевается книга целиком.

В общем виде процесс подготовки разбивается на три фазы, хотя в зависимости от формата издания и целей работы возможны вариации.

На фазе оцифровки производятся серийные процедуры: постраничное сканирование издания, программное распознавание текста, первичное вычитывание текста. Основной результат фазы — единый файл в формате Word, содержащий текстовую информацию издания. По необходимости, на этой же фазе готовятся иллюстрации и рисунки в адекватном качестве

Подобная процедура весьма стереотипна, она подразумевается в любом варианте подготовки электронного издания, не только в обсуждаемом. Выполнение её, безусловно, трудоёмко, но требует от исполнителей достаточно стереотипных навыков. Качество результата этой фазы невысокое, что осложняет технологию подготовки на двух следующих фазах.

Фаза разметки на данный момент производится в тестовом процессоре Word, с использованием специально разработанных средств. Здесь происходит оформление логической структуры издания и включённых в него произведений, а также всех структурных элементов, связанных с

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

текстом: ссылки, сноски, иллюстрации, страницы и их нумерация. Одновременно производится окончательное вычитывание.

Синхронно с разметкой, для всего издания и для произведений оформляются библиографические описания.

На фазе окончательной подготовки издание разделяются на отдельные произведения, оформленные в формате HTML. Здесь производятся проверки целостности информации и комплексной оценки качества результата.

Две завершающие фазы подразумевают от исполнителей значительно более высокий уровень ответственности и специализированной подготовки. Существующие на данный момент программные средства явно недостаточны, чтобы заменить ручной труд этих специалистов без потери качества результата. В то же время, вполне возможно радикально облегчить их труд, автоматизировав наиболее трудоемкие части процедур.

В настоящее время мы начали модернизацию наших технологий, исходя именно из этого принципа: ***создать среду, в которой процессы фаз разметки и окончательной подготовки были бы поддержаны программно — причём максимально удобным и современным способом***. Необходимо заметить: добиваясь эстетических качеств у программного продукта, мы в конечном итоге сокращаем издержки технологии и существенно оптимизируем затраты.

3 Роль среды Word в текущей технологии и в условиях модернизации

Дополнительная цель, преследуемая при создании среды: избавиться от издержек в использовании промежуточного формата Word, и в частности слить фазы разметки и окончательной подготовки. В связи этим уместно сделать следующий комментарий.

В период формирования технологии во второй половине 90-х годов, выбор текстового процессора Word был обоснованным решением. Он позволил автоматизировать многие ручные процедуры, характерные для процесса подготовки. В то же время, концепция электронного документа, поддерживаемая редактором Word, отличается от необходимой для наших целей концепции электронного издания, ориентируемой на возможности формата HTML. Различия эти на первый взгляд незначительны, но в рамках технологии, при подготовке больших текстовых массивов, приводят к очень серьёзным трудностям. Учёт этих трудностей привёл к необходимости поддерживать в среде Word не только систему "скриптов", но и специальный язык разметки документов. Выстроенная таким образом и существующая на данный момент технология оказалась вполне работоспособной, но весьма тяжеловесной.

С момента создания нашей технологии, средства Word по поддержке формата HTML были развиты радикально. Однако с того же времени столь же радикально были развиты и другие средства текстового редактирования, в частности и средства непосредственного редактирования файлов в формате HTML. Поэтому в контексте задач по разметке текста принципиальной необходимости в промежуточном формате Word на нынешний день нет.

В то же время, формат Word остаётся базовым для первого этапа подготовки - оцифровки. Наиболее удобные из наработанных средств для формата Word планируется перенести сюда, частично разгрузив тем самым более ответственную последующую фазу разметки.

4 Анализ доступных решений

В принципе, подготовку изданий в желаемом формате можно выстроить вокруг любого текстового редактора. Однако процесс подготовки включает много различных операций, и для достижения качества результата все эти операции необходимы. Существенная особенность нашей задачи: операции должны проводиться над большими массивами текста. Таким образом, формальный список требований (вернее, пожеланий) к создаваемой среде состоит из многих пунктов, каждый из которых в отдельности не имеет принципиального значения. Общее требование: оптимизировать по возможности большую часть разнообразных операций.

Приступая к построению среды, мы ориентировались на максимальное использование уже готовых программных решений, соответственно, к минимизации собственных затрат на разработку. Анализ доступных решений выявил довольно странную картину.

Оказалось, что существует множество вариантов, которые реализуют требуемую нами функциональность, но только по частям, фрагментарно. В целом, использование каждого из готовых решений для наших целей оказалось не удобно. При этом существенно, что открытое программное обеспечение по функциональности показало себя ничем особенным не хуже и не лучше, чем коммерческие варианты.

В результате мы отказались от варианта взять определённый программный продукт, и выстраивать наше решение на нём одном, пользуясь возможностями «плагинов» и настроек. Вместо этого было принято решение создавать самостоятельную программу, используя доступные решения с открытым кодом (Open Source).

5 Анализ концепции интерфейса

Дополнительный анализ показал, что ключевой частью задачи по построению нашей среды оказывается концепция интерфейса. Технические же средства — поддержка форматов, алгоритмов —

доступны во многих качественно оформленных вариантах, и их использование не составляет тяжёлой проблемы.

К решению проблемы интерфейса нас подтолкнуло следующее обстоятельство, которое можно было бы назвать недоразумением. Общеизвестный стереотип интерфейса под названием «редактор» на самом деле предоставляет средства не для *редактирования* документов, а для их *создания*. Соответственно, при построении нашей среды нам важно реализовывать как раз те части функциональности, которые в стереотипе «редактора» делаются неудобно. И наоборот: важнейшие для стереотипа задачи в нашем случае оказываются второстепенными:

- «в редакторе» главное время пользователя тратится на создание текста; «в среде» - это происходит довольно редко;
- «в редакторе» документ, как правило, маленький — его создают; «в среде» он всегда большой
- «в редакторе» средства навигации не слишком важны, а структуры, по которым производится навигация (ссылки, сноски, разделы) очень динамично изменяются; «в среде» навигация крайне важна, а навигационные структуры требуют постоянной проверки на корректность, а также внятной маркировки проблемных ситуаций
- «в редакторе» запуск «скриптов», делающих автоматические проверки и исправления, производятся в свободном порядке; «среда» должна действительно быть средой, в которой автоматические процедуры могут запускаться на регулярной основе, а отработка ошибок и сообщений этих процедур должна быть удобной и очевидной

6 Программная среда для подготовки электронных документов

Отталкиваясь от этих наблюдений, у нас сложилось следующее интерфейсное решение.

Среда выстраивается вокруг режима просмотра документа и его оглавления. При этом всевозможные ошибки и конфликты, неизбежные при подготовке, отображаются пометками на полях.

Операции редактирования при подготовке разделяются на две группы. Если автоматические процедуры («скрипты») могут выполняться как на всём массиве, так и на его частях, пользовательские операции выполняются только на небольших фрагментах, как правило, размером в один, максимум несколько абзацев текста.

Автоматические процедуры, используемые нами в технологии, очень сильно зависят от контекста конкретной задачи. Как правило, конкретная процедура не может быть рассчитана на «все случаи жизни», и возможность отследить «неправильную ситуацию» при запуске процедур — крайне важная техническая задача. Предлагаемая схема интерфейса дает такую возможность: процедура получает возможность оставить после своего выполнения «заметку на полях», в наглядном и удобном виде.

Пользовательские задачи, требующие повышенного внимания, в данном решении реализуются, как операции над ясно определёнными фрагментами текста. Таким образом, среда гарантирует, что при исправлении фрагмента не испортятся другие фрагменты, и процесс редактирования выстраивается, как понятная последовательность атомарных процедур.

7 Платформа разработки

Рассмотрев различные варианты, в качестве платформы для своего проекта мы выбрали решение из проекта Mozilla. Конкретно, это версия продукта XUL Runner, с включённой поддержкой всех возможностей языка Python. Это решение было стабилизировано разработчиками Mozilla сравнительно недавно, и обещает стать популярным в разнообразных мировых разработках открытого программного обеспечения.

Привлекательность решения, с нашей позиции, заключается в следующем свойстве комбинированного продукта. XUL Runner представляет собой не готовое приложение, а среду разработки приложений, обеспечивающую полноценный доступ к ядру, общему для всех продуктов проекта Mozilla. В рамках самой среды можно создавать различные автономные ("клиентские") приложения с определённым креном в сторону сетевых решений. Это очень удобно для наших задач, поскольку результат нашей технологии — файлы в формате HTML — предназначены для сетевого использования. В то же время, использование только XUL Runner порождает определённые проблемы: встроенные средства языка Java Script вполне достаточны для клиентской части сетевого решения, но представляются недостаточно удобными для полновесного локального приложения, каким является наша среда.

Подключение к конфигурации языка Python со всеми его стандартизированными пакетами полностью снимает этот недостаток, открывая доступ к всевозможным общедоступным техническим средствам.

8 Заключение

Предыдущий вариант нашей технологии разрабатывался с середины 90-х годов прошлого века. За прошедшее время он, безусловно, морально устарел. В то же время, до последнего времени сохранялась ситуация, когда состояние общих программных технологий не позволяло существенно облегчать процесс подготовки текстов. По нашим предположениям, сейчас ситуация изменилась, и доступные в настоящее время

средства достаточны, чтобы эффективно провести модернизацию процесса подготовки.

На момент написания этого доклада обсуждаемая программа (техническое название "H4") находится в завершающей стадии разработки, и готовится к внедрению в технологический процесс.

Литература

[1] Вигурский К.В. К проблеме оценки качества электронных библиотек. *Электронные библиотеки России: управление и координация: Всероссийская научно-практическая конференция*, Москва, РГБ, 21–22 февраля 2007 г.

[2] Вигурский К. В. Представление текстовых произведений в электронной форме (Опыт Фундаментальной электронной библиотеки "Русская литература и фольклор"). *Книга и мировая цивилизация: Материалы XI Международной научной конференции по проблемам книговедения* (Москва, 20–21 апреля 2004 г.): В 4 т. – М.: Наука, 2004. – Т. 1. —С. 346 – 386

[3] Вигурский К. В., Пильщиков И.А. Филология и информатика. *Известия АН*. Серия литературы и языка, 2003, Т. 62, № 2. – С. 9–16

[4] Кузнецов Ф.Ф., Вигурский К. В. Фундаментальная электронная библиотека «Русская литература и фольклор»**.** *Вестник Российского гуманитарного научного фонда*, 2004, № 3 (36). – С. 54 – 63

[5] Creating Python GUI Applications using XULRunner. http://pyxpcomext.mozdev.org/no_wrap/tutorials/pyxulrunner/python_xulrunner_about.html

Technological process for preparation of publications, on example of Fundamental electronic library «Russian literature and Folklore». Current state and the modernization principles.

S.I.Trifonov, A.E.Polyakov

We discuss the technology used for transformation of publications from their printed form to electronic one. This technology was developed and adapted in the Electronic library dept. of the SRC INFORMREGISTR, in particular for the Fundamental electronic library «Russian literature and Folklore» (http://www.feb-web.ru). The current state and the principles of modernization are considered.