ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ ЦНСХБ – ПРОБЛЕМЫ И РЕШЕНИЯ

Аветисов Михаил Андреевич, Центральная научная сельскохозяйственная библиотека, 107139, Москва, Орликов пер., 3В, am@cnshb.ru
Крамчанинов Евгений Викторович, Центральная научная сельскохозяйственная библиотека, 107139, Москва, Орликов пер., 3В, kjv@cnshb.ru
Стеллецкий Василий Игоревич, Всероссийский научно-исследовательский институт информации и технико-экономических исследований агропромышленного комплекса (ВНИИТЭИагропром), 129337, Москва, Хибинский проезд, д.20, sw@cnshb.ru

DIGITAL LIBRARIES CSAL – PROBLEMS AND SOLUTION

Avetisov Michael A., Central Scientific Agricultural Library, 107139, Moscow, Orlikov bystreet, 3B, am@cnshb.ru
Kramchaninov Evgeny V., Central Scientific Agricultural Library, 107139, Moscow, Orlikov bystreet, 3B, kjv@cnshb.ru
Stelletshky Vasily I., Research Institute of Information and Technical Economic Studies in Agrobusiness., 129337, Moscow, Khibinsky proezd, 20. sw@cnshb.ru

Creation subjects of the digital library of the Central Scientific Agricultural Library was discussed in the article. The digital library contents encyclopedic and referencial documents/ Matters of source data structuring, text markup and recognition and database build up was discussed. Query building and using the linguistic information retrieval system «Artefact» was considered.

Работа выполнена при финансовой поддержке РФФИ, грант № 00-07 90208.

Начиная с 1999 года, ЦНСХБ приступила к созданию электронных библиотек. К этому времени стало очевидно, что происходит бурный рост числа электронных библиотек, обеспечивающих пользователям доступ к первоисточникам, как опубликованным на бумаге и имеющих копию в электронном виде, так и появляющимся только в электронном виде. При этом в подавляющем числе случаев держателями этих библиотек становятся издательства, консорциумы издательств и подобные организации, являющиеся генераторами информации. Что касается традиционных библиотек, развивающих свои электронные ресурсы, то они обычно формируют свои электронные библиотеки на основе различных произведений, отсканированных или полученных на электронных носителях. Обычно это тексты или изображения оригинальных материалов. Поиск информации обычно осуществляется на уровне целого документа по его описанию

(библиографическому или на классификационных языках), а навигация в лучшем случае по оглавлению [1].

Имеется и еще одна ветвь электронных библиотек, представляющих энциклопедическую или словарную информацию. Появление подобных электронных ресурсов совпало по времени с созданием и развитием наших электронных библиотек. Наиболее интересным электронным ресурсом является сайт «Рубрикон» (www.rubricon.ru), повлиявший на выбор репертуара выставляемых на нашем сайте энциклопедий и справочников. Предполагаемый «репертуар» нашей электронной библиотеки изложен в [2].

Подход разработчиков к созданию электронной справочной библиотеки, которую мы условно назвали «сельскохозяйственная электронная библиотека знаний» (СЭБиЗ) состоит в следующем. Осуществляется построение информационного ресурса справочно-энциклопедического характера по основным наиболее важным вопросам сельскохозяйственной науки и практики. В качестве источников информации при подготовке ресурса используются наиболее значимые справочные и энциклопедические издания, которые, особенно теперь, труднодоступны в большинстве регионов. Появление «Рубрикона» (сельскохозяйственная и ветеринарная энциклопедия в СЭБиЗ появились раньше) с его мощными ресурсами позволит нам в дальнейшем создавать только специализированные справочники и отойти от энциклопедий.

Исходный материал

При создании СЭБиЗ (http://www.cnshb.ru/akdil) нами было проанализировано достаточно большое количество энциклопедий и справочников, что позволило разработать основные алгоритмы структурирования полных документов на статьи и обеспечить естественную навигацию по справочникам. Практически все подобные документы представляют собой иерархические структуры не более 4-х уровней иерархии. Обычно в тексте имеются перекрестные ссылки, при этом, зачастую, не в именительном падеже, в ряде случаев и с нарушением порядка слов. В текстах встречается значительное число таблиц, формул и изображений. Таблицы и изображения зачастую привязаны к определенному месту из-за полиграфических требований, и не относиться к тому фрагменту справочника, в котором они встретились. Для выделения статей и ссылок обычно используются шрифтовые выделения, абзацные отступы, специальные слова типа «см», «см. также», «рис», и т.п.

На основе анализа разработана система разметки текстов исходных документов, позволяющая создать правильную структуру электронного справочника. Разметка существенно усложняет корректурный процесс, требует от технического персонала понимания структуры справочников и удорожает процесс подготовки их исходных данных. Разработчики пошли

по пути минимизации разметки текста и максимального использования структурных выделений в тексте исходного документа. (Это особенно будет важно в дальнейшем при формировании новых разделов библиотеки из средств ЦНСХБ). Поэтому разработана не только программа обработки размеченного текста, но и специализированный набор модулей, позволяющий легко подстраивать программные средства под каждый конкретный справочник.

Технология создания документа

Сканирование документов осуществляется на высокопроизводительном книжном сканере Bookeye. Качество текстового и графического штрихового изобразительного материала достаточно высокое для последующего распознавания. Остальной изобразительный материал сканируется на планшетном сканере, а при невозможности – фотографируется цифровым фотоаппаратом. В этом случае связь картинок с соответствующими страницами, отсканированными на книжном сканере, обеспечивается выбором системы именования папок, файлов и номеров рисунков.

Распознавание осуществляется программой FineReader v.5.0 и v.6.0. Корректура и необходимая разметка текста осуществляется непосредственно в этой программе. Выгрузка в формат редактора Word осуществляется с сохранением шрифтовых выделений оригинала, что позволяет в дальнейшем осуществлять на этой основе структурирование документа. Корректировка может производиться и в редакторе Word, что позволяет осуществлять корректурный процесс вне стен библиотеки. В этом же редакторе осуществляется набор формул (оставленных в виде картинки после распознавания). Формула (объект с точки зрения Word'а) специальным образом помечается, так же как и подписи под рисунками. Это позволяет в дальнейшем обеспечить обтекание текста только вокруг картинок, а не формул, и правильно расставлять подписи под рисунками. Откорректированный и частично (при необходимости) размеченный текст переводится средствами Word'а в HTML формат.

Последующая обработка осуществляется специальными программными средствами, разработанными для формирования ЭБ. Программы написаны на языках C++ и REFAL и включает специализированные модули обработки тех или иных особенностей текста. В процессе обработки осуществляется структуризация массива данных, разбивка на отдельны документы, обработка разметки в соответствии с видом документа, обработка стандартных элементов разметки, составление словарей ссылок с учетом орфографии русского языка (в основном это анализ окончаний), разрешение ссылок, формирование ссылок на отдельные рисунки и таблицы, осуществляется разметка документа для загрузки в базу данных ИПС, т.е. простановка меток полей и т.п. Подготовленные документы в формате

HTML загружаются в базу информационно-поисковой системы «Артефакт». Каждый справочник или энциклопедия представляет собой отдельную базу данных.

База данных и поисковые возможности

Мы исходили из того положения, что при обращении к электронной библиотеке пользователь в большинстве случаев будет осуществлять тематический поиск на полных текстах справочного материала. Поскольку библиотека задумывалась как справочно-энциклопедическая, то, по-видимому, значительная часть обращений будет представлять собой поиск термина или устойчивых словосочетаний. Кроме того, следует отметить, что контингент пользователей в сельскохозяйственной науке и практике имеет недостаточную компьютерную грамотность и опыт работы с электронными информационными ресурсами. Из всего сказанного следует, что весьма желательно, чтобы запрос можно было бы составлять на достаточно простом языке, лучше, близкому к естественному.

Поисковые механизмы электронной библиотеки реализованы на средствах информационно-поисковой системы «Артефакт», как наиболее удовлетворяющей нашим требованиям. Язык запросов ИПС Артефакт включает в себя традиционные операторы И, ИЛИ, НЕ, операторы близости слов, наличия слов в одном предложении, операторы обработки элементов типа «дата» и т.п. Однако задание запроса требует знания синтаксиса языка. В связи с этим был разработан синтаксический анализатор запроса. При его использовании пользователь может составлять запрос практически в виде обычного текста, при необходимости используя или не использовать скобки для исключения неоднозначности. Морфологический разбор слов поисковой системой позволяет не задумываться о падежах, склонениях и спряжениях.

Развитие синтаксического анализатора предполагает возможность расширения запроса за счет подключения словарей (например, латинских названий растений) и/или сельскохозяйственного тезауруса, который в настоящее время используется для расширения описания документов электронного каталога ЦНСХБ.

В настоящее время осуществляются работы по подключению тезауруса для расширения запроса, что может быть очень полезно при поиске по полным текстам.

Каждый справочник представляет собой иерархическую структуру документов в формате HTML, которые загружаются в базу данных ИПС. Следующие версии ИПС смогут работать и с документами в формате XML. Поиск может осуществляться как по отдельным базам данных, так и по всем базам одновременно.

Разработчики ЭБ исходили из того, что библиотека подобного рода будет весьма полезна сообществу пользователей информации в области сельского хозяйства. Поэтому предполагается, что работы по наполнению библиотеки будут продолжаться и в дальнейшем (по окончанию гранта), но, что достаточно очевидно в условиях Россельхозакадемии, при незначительном финансировании. Это, в свою очередь, может повлечь увеличение ошибок из-за снижения качества корректурного процесса, как самого дорогого во всем производственном процессе создания ЭБ. В связи с этим ЭБ включает в свой состав базу данных обратной связи. База создана под управлением СУБД MS SQL v.7.0. При просмотре любого документа пользователи имеют возможность написать о замеченных ошибках. Эти тексты заносятся в записи СУБД, соотнесенные с документом справочника. Разработана программа-менеджер работы с базой обратной связи, позволяющая администратору наблюдать появление замечаний для всех баз данных ЭБ, просматривать их, а также просматривать и редактировать соответствующие документы электронной библиотеки. Редактирование осуществляется одним из редакторов текстов HTML-формата, обеспечивающего возможность просмотра и редактирования документа также и в текстовом формате.

В состав электронной библиотеки могут входить «старые» документы, т.е. документы, представленные в дореформенной (1918 года) орфографии. Мы включили в электронную библиотеку книгу С.П. Урусова «Книга о лошади» с тем, чтобы отработать некоторые особенности представления такого рода литературы в СЭБиЗ. Книга структурирована на основе оглавления, каждая страница – отдельный документ. В отличие от обычных справочных материалов, документы кроме перекрестных ссылок, если таковые появляются в тексте, имеют две дополнительные ссылки – на следующий или предыдущий документ (страницу) и ссылку на «оригинал» (в PDF-формате). Распознавание осуществлялось стандартными средствами (FineReader) с обучением графике и орфографии. Как мы и предполагали, к настоящему времени FineReader почти «научился» квалифицированно работать с дореформенной орфографией. Приведение текста к современному алфавиту и орфографии позволяет пользователю искать необходимые ему фрагменты текста, термины слова и факты на современном языке с последующим просмотром оригинала.

В состав библиотеки входят и справочные материалы по научно-исследовательским организациям, опытным станциям, учебным заведениям агропромышленного комплекса. В настоящее время это структурированный набор страниц Web-сервера, включаемый также в базу данных ИПС «Артефакт». Разработчики отдают себе полный отчет, что соответствующая база данных должна строиться на базе языка XML с использованием соответствующих метаданных, примером чему может служить система ИСИР РАН. ЦНСХБ участвует в проекте LibWeb2, также поддержан-

ного РФФИ, в рамках которого предполагается создание схем и правил разметки, а также необходимых программных средств для подобного рода данных. Мы надеемся, что это позволит нам провести реструктуризацию уже подготовленного и выставленного в рамках ЭБ материала по организациям, персоналиям и научным работам этих учреждений.

Следует упомянуть и электронную библиотеку отчетов по научно-исследовательским работам, которая создавалась в 1998-2000 годах также при участии авторов. Эта работа позволила отработать механизм разметки отчетов, необходимый для представления отчета в виде структурированного документа, а также механизм связывания рисунков. В настоящее время эта библиотека пополняется новыми отчетами. Электронная библиотека выставлена в Интранет и предоставляется читателям библиотеки только в стенах библиотеки в соответствии с решением Министерства сельского хозяйства, собственника информационного ресурса.

Литература

1. Груздев И.А., Лавренова О.А., Перли Б.С. Электронные библиотеки РГБ – составная часть РГБ // Третья Всероссийская конференция по электронным библиотеками, Петрозаводск, 11–13.09.2001
2. Аветисов М.А., Крамчанинов Е.В., Стеллецкий В.И. Новое направление в информационном обеспечении сельскохозяйственной науки // Новые технологии в информационном обеспечении науки, М., Биоинформресурс, 2001