Автоматизированная система аналитической росписи документов

The Automated System of an Analytical List of Documents

Автоматизирована система аналітичної розписі документів

Аветисов М. А.
Центральная научная сельскохозяйственная библиотека (ЦНСХБ), Москва, Россия

M. A. Avetisov
Central Scientific Agricultural Library (CSAL), Moscow, Russia

Аветісов М. А.
Центральна наукова сільськогосподарська бібліотека (ЦНСГБ), Москва, Росія

Рассматривается проблема сокращения ручного ввода данных при описании статей из журналов и научно-технических сборников для электронного каталога ЦНСХБ.

The problem of reduction of manual data input is considered at the description of articles from magazines and scientific and technical collections for electronic catalogue CSAL.

Разглядається проблема скорочення ручного ведения данних при опису статей із журналів і науково-технічних збірок для електронного каталогу ЦНСГБ.

Создание записей для электронного каталога является весьма трудоемким делом. Поэтому создаются всевозможные объединения и консорциумы для разделения труда при создания библиографических записей. Однако это касается в основном записей, описывающих отдельное издание — книгу, сборник, отдельный номер журнала и т. п. или коллекцию — многотомник, журнал в целом и т. п. Библиотек, осуществляющих постатейную роспись журналов и сборников в достаточно большом объеме, совсем не много.

Еще в конце 80-х годов наша библиотека совместно с ВНИИТЭИсх (теперь уже не существующему) приступила к аналитической росписи ядерных сельскохозяйственных журналов. А с 1992 года, создав 1-ую версию своей автоматизированной системы, обеспечивает ввод данных в свой электронный каталог (ЭК) описаний статей. При этом ЭК статей на порядок больше по объему ЭК книг. Поэтому объем данных, вводимых только о статье из журналов или сборников, существенно больше, чем объем описания отдельных изданий. Если можно уменьшить трудоемкость ввода этих данных, то на лицо сокращение ручного труда.

Каковы предпосылки создания автоматизированной системы аналитической росписи документов.

1. В качестве основного материала для ввода данных о статье рассматриваются оглавления журналов или сборников. Следует обратить внимание и на следующий факт. Полнотекстовые электронные ресурсы, особенно иностранные журналы, стали занимать значительное место в информационном обеспечении пользователей библиотеки. И всегда имеется достаточно формализованное оглавление.
2. Оглавление в научно-технических журналах и сборниках обычно легко выделяется среди остального текста (что нельзя сказать про огромное количество журналов массовой культуры)
3. Заголовки статей, особенно в сборниках, несут в большинстве случаев, информацию о содержании статьи. Кроме того, научные журналы и сборники зачастую помещают статьи в рубрики или разделы, которые также отражаются в оглавлении.
4. Структура строк оглавления, описывающих каждую статью, для каждого журнала стабильна, как минимум в течение года. Практически существует всего несколько видов структур, типа «Авторы, Название, Страницы» или «Название, Авторы, Страницы» и т. п.
5. ЦНСХБ имеет развитый тезаурус по агропромышленному комплексу и пищевой промышленности, что позволяет сопоставлять термины тезауруса со словами и словосочетания из названия и обеспечивать обогащение описания статьи.
6. Существующие оборудование сканирования (книжные сканеры, имеющиеся в ЦНСХБ) позволяют осуществлять сканирование с высокой скоростью. Сканированию подвергаются новые, поступающие в библиотеку журналы и сборники, поэтому качество сканирования высокое.
7. Возможность автоматического или полуавтоматического мониторинга подписанных зарубежных баз данных (например, Agricola или отдельных баз данных EBSCO) и создания копий соответствующих оглавлений на собственном Web-сайте.
8. Мы полагаем, что пользователь библиотеки ищет в основном либо по автору и/или названию статьи источника или же по тематике. В последнем случае пользователя мало интересует (с

некоторой точностью, не существенной в данном случае), в каких полях встречаются важные для него термины или рубрики. И большинство сайтов так и устроены.

Все эти предпосылки побудили нас к созданию автоматизированной системы аналитической росписи документов. Можно рассматривать ее как набор отдельных подсистем:

a. подсистема формирования оглавлений журналов и сборников;
b. подсистема дополнительной ручной обработки описания;
c. подсистема автоматического формирования записи для ЭК;
d. подсистема учета выполнения исполнителями всех операций над данными в АСАРД

Подсистема a) имеет два модуля, один из них — модуль сканирования и распознавания печатных изданий. В подавляющем большинстве случаев распознанный и проверенный текст не требует дополнительной ручной доработки. В отдельных случаях используется специальный язык разметки. Текст сохраняется в HTML-формате, а образ, который также доступен пользователю, просматривающему оглавление, в PDF-формате.

Второй модуль — это модуль обработки оглавлений on-line (т. е. внешних баз данных). В частности, для БД Agricola обеспечивается мониторинг оглавлений и подкачка оглавлений вновь появившихся номеров на сайт ЦНСХБ.

Подсистема b) позволяет пользователю разметить оглавление и передать отдельные статьи не обработку автомату или конкретному исполнителю. Оглавление автоматически разбирается по полям, исполнителям с разными правами доступа (каталогизатор, систематизатор, распределитель работы и т. п.) предоставляется возможность работы только со своей группой полей как заполненных автоматических, так и дополнительными. Кроме того можно проставить признак готовности документа, просмотреть историю работы с ним. Поскольку ЦНСХБ выпускает еще и реферативный журнал, то можно приписать реферат, направить документ в тот или иной выпуск (Ветеринария, Пищевая промышленность и т. п.) и номер.

При вводе данных обеспечивается интерактивный контроль орфографии. Возможна также проверка правильности подготовки всей записи средствами ОРФО, а также обогащение словаря терминов для последующего анализа и обогащения тезауруса или специальных словарей.

Подсистема c) обеспечивает создание записи электронного каталога статей на основе подготовленных автоматически или с добавлением данных ручного ввода документов.

Название статьи разбирается на слова. Не рассматриваются стоп-слова и цифро-буквенные слова. Все слова нормализуются программой А. Сокирко (http://www.aot.ru/technology.html) с необходимыми для наших целей доработками (в дальнейшем средства могут быть и другими, — нет предел совершенствованию).

Слова сравниваются с терминами тезауруса на предмет близости. Наиболее близкие термины заносятся в поле «термины тезауруса». Поле «Рубрики ОРНТИ» заполняется рубриками из общего описания журнала и/или рубриками, связанными с терминами тезауруса.

Предполагается создание специального словаря на основе текстов постоянных или часто встречающихся рубрик или разделов оглавлений. Этот словарь может быть обогащен кодами рубрик ОРНТИ. В этом случае эти коды будут включены в соответствующие элементы данных.

Предполагается расширение тезауруса переводами терминов на английский язык (или на другие языки). В этом случае, при обработке иностранных названий статей возможно включение русских дескрипторов в соответствующее поле в дополнение к англоязычным. Это обеспечит необходимые «подсказки» для пользователей при поиске информации.

Следует заметить, что как только появляется оглавление журнала или сборника в электронной форме, то вслед за эти появляется и запись электронного каталога. Ручное обогащение записи или изменение ее — асинхронный процесс, с неопределенным временем окончания. Пользователь электронного каталога будет видеть улучшающуюся во времени запись по мере того, как специалист будет приписывать (или, при необходимости, исправлять) соответствующую запись с АСАРДе, правда с недельным опозданием (периодичность обновления данных ИПС).

Подготовленные записи загружаются в информационно-поисковую систему АРТЕФАКТ (разработка ИА «Интегрум-техно»).

Подсистема d) обеспечивает учет всех операций с записью, которые осуществляет пользователь. Он входит в систему со своим идентификатором. В отдельных случаях, например, при поручении работы конкретному исполнителю, идентификатор этого исполнителя вводится в систему дополнительно. По всем операциям запоминается дата, время, объем изменений, идентификатор исполнителя, а в ряде случаев и другая необходимая информация. Это позволяет наладить полный компьютерный учет работы коллектива исполнителей.

Дополнительно следует заметить, что ЦНСХБ имеет электронный архив документов, в котором хранятся образы оглавлений (если они есть), а также все полные тексты статей, полученных как в результате

работы службы ЭДД (заказ из ЭК ЦНСХБ) так и тексты всех статей, «скаченных» сотрудниками ЦНСХБ из внешних баз данных. Все полные тексты «привязаны» к соответствующим записям ЭК и оглавлениям журнала. Внутри библиотеки или с удаленных терминалов (в других организациях, технология VPN-соединений) они доступны для чтения, а для пользователей сети Интернет — для заказа.

Система полностью базируется на СУБД MS SQL. Для каталогизации сборников, как отдельных изданий, и журналов в целом используется система Марк-SQL (Информсистема), так как их библиографическое описание сложно и требует специализированного программного обеспечения.

Каждый документ, поступающий в ЦНСХБ, снабжается электронным номером (ЭН), напечатанным на наклеивающейся этикетке и отображаемым в виде штрихкода. ЭН является полем связи для объединения различных видов описания документа и его частей.