

## **Сельскохозяйственная электронная библиотека знаний: проблемы и решения**

***Аббакумова Н.П., Аветисов М.А.***

**ЦНСХБ Россельхозакадемии**

---

В настоящее время творческим коллективом сотрудников ЦНСХБ Россельхозакадемии в рамках гранта РФФИ № 00-07-90208, проводится целый комплекс мероприятий по созданию Сельскохозяйственной электронной библиотеки знаний (СЭБиЗ, <http://www.cnsnb.ru/akdil>).

Разработчики поставили перед собой цель - определить политику формирования электронной сельскохозяйственной библиотеки знаний и обеспечение доступа к ней на основе новых информационных технологий в сочетании с простыми средствами взаимодействия с пользователем на языке, близком к естественному, профиль ее комплектования, приоритеты при отборе документов, включая электронную версию фрагментов редкой сельскохозяйственной книги, а также уже имеющихся в ЦНСХБ электронных информационных ресурсов. Создание сельскохозяйственной электронной библиотеки знаний (СЭБиЗ) предполагает принципиально новый подход к проблеме кумуляции информации с использованием новейших информационных технологий с целью обеспечения оперативного доступа к справочной информации ученых и специалистов агропромышленной отрасли.

ЦНСХБ Россельхозакадемии как отраслевой информационный центр создает целый комплекс вторичной информации по проблемам АПК, который включает: текущую систематическую библиографическую информацию, реферативную, библиографическую информацию научно-вспомогательного характера, биобиблиографическую.

Поставив перед собой задачу создания СЭБиЗ, была выработана основа ее содержания, которая включает в себя:

- фрагменты энциклопедических словарей и справочников.
- сведения о научно-технических разработках;
- полнотекстовые фрагменты из редкой книги;
- справочно-адресную информацию о научно-исследовательских и учебных заведениях АПК;
- рисунки, таблицы, графики, фотоматериалы;
- полнотекстовые отчеты по НИР.

Значительные усилия в процессе выполнения проекта были направлены на выявление информационных ресурсов по научному обеспечению сельского

хозяйства в нашей стране и за рубежом.

До сегодняшнего времени наиболее приемлемой формой представления информации по проблемам АПК являются базы данных, записанные на электронных носителях, либо выставленные в ИНТЕРНЕТ. Как правило, они основаны на результатах лингвистической обработки документного потока научных библиотек как российских, так и зарубежных, а также различных международных информационных организаций в сфере аграрной отрасли, в частности: FAO - Food and Agriculture Organization of the United Nations, IAALD - International Association of Agricultural Information Specialists and CAB International. Создан целый ряд мировых сельскохозяйственных узлов, где пользователь может получить информацию об электронных информационных ресурсах по интересующей его проблеме. К сожалению, в России практически не существует узлов, на которых отражались бы научные достижения организаций и предприятий сельскохозяйственного профиля. Имеют место случаи предоставления пользователям информации по инженерно-техническому обеспечению АПК России, удобрениям, кормам или ветеринарным препаратам.

В качестве базового программного обеспечения СЭБиЗ выбраны: MS SQL, ИПС "Артефакт" (разработка информационного агентства Интегрум-Техно), а также программные средства "Электронного архива (ЭА) ЦНСХБ" (разработка Cognitive Technologies). ЭА предназначен для комплексного решения задач ввода, обработки, централизованного хранения и отображения книг, журналов и других печатных документов, а также для хранения и поиска информации, представленной в виде электронных документов (тексты, графические образы, мультимедийная информация).

Назначением системы является организация системы ввода, хранения и доступа к электронным копиям первоисточников, образующим электронный архив, при этом решаются следующие задачи:

- ввод (включая потоковый ввод) и редактирование данных, полученных путем сканирования документов;
- ввод и редактирование данных, полученных из автоматизированной библиотечно-информационной системы ЦНСХБ (АБИС);
- реквизитная и/или полнотекстовая индексация всех хранимых материалов, включая пакетный режим индексирования;
- хранение и предоставление доступа в локальном режиме и в сети Интернет к материалам, составляющим ЭА;
- разграничение доступа для разных категорий пользователей;
- вывод данных по запросу для электронной доставки копий первоисточников пользователям.

Информация для загрузки в ИПС и последующего отображения представлена в формате HTML и включает тексты, рисунки, фотоматериалы, аудио материалы.

Полные тексты документов структурируются и обеспечиваются гиперссылками.

СЭБиЗ формируется как совокупность баз данных. Разработчики предполагают, что динамика изменений библиотеки (т.е. изменения в уже подготовленных

документах) происходит редко, поэтому для обеспечения полноценного и быстрого поиска все данные преобразуются в базы данных ИПС Артефакт. Эта система характеризуется следующими основными параметрами:

- система рассчитана на работу с полнотекстовыми базами данных;
- объем каждого документа практически не ограничен;
- объем баз также практически не ограничен. При этом время поиска мало на гигабайтных объемах баз;
- поиск может вестись одновременно по многим данным;
- поиск осуществляется с учетом морфологии русского и английского языков. В поисковом предписании, возможно, использовать, операторы И, ИЛИ, НЕТ, скобочные конструкции, правостороннее усечения, операторы соседства и порядка слов в предложении;
- имеются операторы поиска по данным типа "дата".

Проект предусматривает интеграцию информационных ресурсов ЦНСХБ (электронных каталогов книг, журналов, статей, авторефератов и т.п., а также полнотекстовой базы отчетов о завершенных научно-технических разработках) в единый интерфейс СЭБиЗ. Разработаны необходимые модули и интерфейсы.

В связи с разнообразием структуры документов: энциклопедии, справочники и словари; книги в дореформенной (1918 г.) орфографии; книги и журналы современной орфографии, были разработаны технологии ввода и разметки документов. Технология предусматривает минимизацию специальной разметки и максимальное использование имеющихся шрифтовых и других выделений, используемых в печатных изданиях.

Разработан механизм и программные средства для формирования на основе отсканированного, распознанного и проверенного (с формированием словаря незнакомых терминов и слов) текста документов по каждой статье словаря с необходимыми гипертекстовыми ссылками, для объединения отдельных фрагментов документа и его частей (текста, рисунков и других графических материалов), подготавливаемых различными исполнителями на разных устройствах сканирования, в один документ с последующей загрузкой в базу данных.

Распознавание текстов осуществляется системой FineReader (v.5.0, v.6.0), которая, в частности, для документов дореформенной орфографии позволяет провести обучение системы распознаванию букв "ять", "и десятичное", "фита" или же имеет соответствующий шрифт, а также слов в дореформенных грамматических формах. Это обеспечивает создание полнотекстовых документов в современной орфографии для качественной работы поисковой системы.

Поиск обеспечивается по полному тексту документов, при этом пользователь может на основании полученного результат просмотреть оригинальный текст в формате PDF или другом (tiff, jpeg). В СЭБиЗ вошел раздел среди информационных ресурсов "Новости сельского хозяйства".

В заключение отметим, что создание СЭБиЗ - это создание справочного, полного, легко обозреваемого, систематизированного и структурированного

информационного продукта с широкими поисковыми возможностями, обеспечивающим ученых и специалистов в области АПК, необходимым набором справочных данных, сведениями о наиболее актуальных проблемах и основополагающих документов в конкретной области знания.

В выводах приведем наиболее общие методические подходы к созданию СЭБиЗ:

1. В основу создания Электронной сельскохозяйственной библиотеки знаний положен принцип INTERNET/INTRANET-технологий; информация для загрузки в ИПС и последующего отображения должна быть представлена в формате HTML и включать тексты, рисунки, фотоматериалы. Полные тексты документов структурируются и обеспечиваются гиперссылками;
2. СЭБиЗ будет представлять собой узел, обеспечивающий доступ к собственным ресурсам по конкретной области знания;
3. Механизм доступа и навигации к информационным ресурсам (ИР) ЦНСХБ Россельхозакадемии должен обеспечиваться за счет соответствующего построения Web-сервера и баз данных.
4. В основу структурирования СЭБиЗ положена конкретная схема ее построения, созданная на основе отобранных документов по определенной области знаний, находящихся в логической связи.
5. СЭБиЗ формируется как совокупность баз данных, для обеспечения полноценного и быстрого поиска все данные преобразуются в базы данных ИПС Артефакт. Поиск должен производиться по всем реквизитам и/или по полному распознанному тексту документа. В поисковом предписании, возможно, использовать, операторы И, ИЛИ, НЕТ, скобочные конструкции, правостороннее усечения, операторы соседства и порядка слов в предложении.

Использование логических взаимосвязей с различными web-узлами позволит расширить информативность библиотеки знаний.

---

### **Об авторах**

**Аббакумова Н.П.** - к.п.н., ученый секретарь ЦНСХБ Россельхозакадемии

**Аветисов М.А.** - зам. директора ЦНСХБ Россельхозакадемии

Государственное научное учреждение Центральная научная сельскохозяйственная библиотека Российской академии сельскохозяйственных наук  
(ЦНСХБ Россельхозакадемии)

Адрес: 107139 Москва Орликов пер.3 кор.В

Тел.: (095) 204-4904, 975-4236

Факс: (095) 207-8972

E-mail: [dir@cnsnb.ru](mailto:dir@cnsnb.ru)

<http://www.cnsnb.ru>

---

© Аббакумова Н.П., Аветисов М.А., 2002