Тезаурус в системе ИРБИС как эффективное средство поиска в ресурсах Корпоративной сети общедоступных библиотек Санкт-Петербурга

Автор: И. Е. Прозоров, М. Н. Сухарева

УДК 025.43

Последовательно отражено решение проблемы представления тезауруса по ГОСТу 7.25-2001 в Системе автоматизации библиотек ИРБИС. Отмечен вклад специалистов ГПНТБ России в решение проблемы.

Ключевые слова: Корпоративная сеть общедоступных библиотек Санкт-Петербурга, информационно-поисковые языки, тезаурусы, словарные статьи, Система автоматизации библиотек ИРБИС.

Корпоративная сеть общедоступных библиотек Санкт-Петербурга (КСОБ) объединяет информационные и кадровые ресурсы около 200 библиотек. Основная её задача - постоянное повышение качества доступа пользователей к совокупным распределённым ресурсам публичных библиотек.

Принцип единой точки доступа для пользователя реализован посредством портала КСОБ (URL: http://ksob.spb.ru/). Портал обеспечивает поиск в сводных каталогах публичных библиотек города, библиографических и полнотекстовых ("Дайджест петербургской прессы") БД. Доступные на портале каталоги и БД раскрывают содержание фондов библиотек - участниц КСОБ, что позволяет оперативно обеспечивать общегородские сервисы МБА, бронирования изданий, ЭДД на качественно более высоком уровне.

Таким образом, поисковые возможности библиографических ресурсов определяют осведомлённость пользователей о составе фондов и задают условия реализации перечисленных сервисов.

Качество поиска в информационно-поисковых системах зависит от множества факторов, в их числе:

уровень информационной культуры пользователей;

степень детализации раскрытия содержания документов;

диапазон поисковых средств;

дружественность используемых информационно-поисковых языков (ИПЯ)

программные возможности АБИС по обеспечению процессов ведения, индексирования и поискового использования конкретных ИПЯ.

Очевидно, что применяемый в библиотечных автоматизированных сис-

темах ИПЯ должен предоставлять малоподготовленным пользователям возможность работы с достаточным по полноте и точности результатом, т.е. быть дружественным. При этом ИПЯ должен обеспечивать достаточно полное раскрытие содержания документа (см. общие требования к ИПЯ [**1**]).

Каждый из применяемых в библиотечной практике ИПЯ имеет свои достоинства. Но наиболее дружественным можно считать тезаурус. По комбинационным возможностям построения поискового выражения он похож на словарь ключевых слов, когда из слов естественного языка в любой последовательности задаются необходимые поисковые признаки. Но тезаурус - нормативный словарь, в котором зафиксированы однозначность и смысловые взаимосвязи терминов. Поэтому при поиске мы достигаем хороших показателей полноты, точности результатов и объёма полезной информации. Сразу уточним: если поиск проводится не наугад, а по терминам поискового словаря. А если АБИС позволяет адекватно отражать на экране всё многообразие связей между терминами (что требует ГОСТ 7.25, п. 4.11.3 [**2**]), то в руках у пользователя оказывается удобный понятийный навигатор.

Отметим, что термины тезауруса позволяют с необходимой детализацией представить все значимые аспекты содержания документа набором самостоятельных точек доступа. Поисковый образ документа строится как дом из множества терминов-"кирпичиков".

Существенное влияние на реализацию конкретного ИПЯ и принятой методики индексирования способно оказывать используемое программное обеспечение. Единой программной средой для всей КСОБ является Система автоматизации библиотек ИРБИС. Библиотеки связаны одной сетью; ищут, вводят и редактируют данные посредством многопользовательского доступа к базам данных на сервере КСОБ. Принятые методические решения, вновь введённая информация сразу же используются в работе всей библиотечной сети.

Аналитическая библиографическая БД Центральной городской публичной библиотеки им. В. В. Маяковского и её поисковый тезаурус ведутся с 1994 г. Сегодня это один из значимых ресурсов КСОБ, создаваемый усилиями десятков библиографов города. Переход нашей библиотеки на ИРБИС в 2001 г. сделал наши каталоги и базы данных доступными читателям для самостоятельной работы внутри библиотеки, а чуть позже - в удалённом режиме. Но отсутствие в системе ИРБИС модуля для работы с нормативными словарями в то время не позволило полноценно реализовать нашу технологию ведения БД: работа с тезаурусом продолжалась в прежней программе CDS/ISIS/M в MS DOS, откуда поисковый словарь выводился в текстовый файл, затем библиографы по алгоритму "найти, выде-

лить, скопировать, вставить" переносили нужный термин в поле ненормированных ключевых слов, а потом - в специальное поле "Дескрипторы".

В 2006 г. после нескольких попыток тезаурус (более 7 тыс. терминов) был перенесён в оболочку базы данных TEZ. Тем самым мы смогли автоматизировать процесс ввода терминов в поле 965 "Дескрипторы" библиографической БД; кроме того, тезаурус стал доступен читателям при "поиске для умников": мы смогли, наконец, представить поисковый словарь в форме, приближенной к тезаурусу, т.е. реализовать не алфавитный, а по-настоящему смысловой отбор поисковых терминов.

Однако процесс индексирования - это ещё не всё. Собственно ведение тезауруса осталось прежним; каждая новая партия терминов из эталонного тезауруса в ISIS переносилась в базу TEZ системы ИРБИС. По-прежнему выводился текстовый файл: он отражал тезаурус во всей полноте связей и использовался библиографами при создании библиографических записей для справок, когда встроенный в 965-е поле словарь не позволял найти "запрятанного" в нём нужного термина.

Функциональные ограничения и отсутствие необходимых полей в рабочем листе базы TEZ создавали существенные проблемы:

жёсткая иерархическая последовательность построения "дерева" тезауруса в TEZ противоречила реальной структуре и самой идеологии тезауруса, созданного по нормам ГОСТа; многие термины были очень условно отнесены к определённой группе; в эталонном тезаурусе большая группа терминов не имела вышестоящих по отношению к себе понятий;

поле вышестоящего дескриптора было неповторяющимся, а ведь достаточно часто термину соответствуют два-три вышестоящих дескриптора: дачи - жилые дома, здания, недвижимость; денежные реформы - валютная политика, денежно-кредитная политика, экономические реформы и др.;

отсутствовало поле для нижестоящих терминов: они отображались в окне просмотра (поскольку в структуре словаря присутствовал вышестоящий термин к каждому из них), но ведение тезауруса должно предусматривать процессы его редактирования, в том числе, вызванные изменениями формулировок и статуса терминов (дескриптор/недескриптор);

отсутствовали поля для систематических и категориальных индексов; эти признаки позволяли редактору тезауруса при работе с базой поискового словаря выводить термины одной области знания, одной категории, чтобы выявить неточности или ранее принятые методические решения в отношении определенных групп терминов; и др.

Результатом серии встреч с главным программистом ИРБИС А. И. Бродовским (ГПНТБ России) в конце 2010 - начале 2011 г. стала реализация нашего тезауруса в оболочке БД URUB, специально предназначен-

ной для разработки различных лексикографических БД. Адекватное и полное отображение связей между терминами было буквально "нарисовано" средствами гипертекстовой разметки (html) в ИРБИС-навигаторе.

Были удачно дифференцированы три категории ссылок (синонимичных связей) от терминов-недескрипторов: "см" (отсылка к одному термину - полному синониму), "см. альтернативу" (отсылка к нескольким из предлагающихся на выбор терминам), "см. комбинацию" (отсылка к двум-трём терминам, создающим в совокупности необходимое понятие).

Решение основных проблем отображения поискового словаря и полноценной навигации в нём средствами гиперссылок как при создании библиографических записей, так и в режиме "поиска для умников", качественно изменило всю нашу работу. Кроме того, это решение сразу вошло в производственный процесс создания БД и информационного обслуживания жителей мегаполиса. Результаты представлены на конференции "Крым-2011" и опубликованы [**3**], а также кратко освещены на форуме системы ИРБИС (URL: http://irbis.gpntb.ru/read.php?10,46024).

Но в ходе годичной эксплуатации тезауруса выявились существенные (хотя и частные) препятствия для полного перехода ведения тезауруса в системе ИРБИС.

Дескрипторная (словарная) статья тезауруса обычно содержит (отображает) по определенной ГОСТом последовательности следующие группы сведений: один-три вышестоящих термина; термины-недескрипторы (синонимы), от которых в самостоятельных ссылочных записях тезауруса дается ссылка к заглавному термину; лексическое примечание (дефиниция термина или методическая рекомендация); нижестоящие термины (до 100 штук: народы - абхазы, японцы); ассоциативные термины (до 30 тематически близких терминов иной, чем заглавный термин, категории: народы - кочевые племена, национальный характер, этнологические исследования и др.).

Хорошо себя зарекомендовавший ИРБИС-навигатор в html выводил термины дескрипторной статьи в рамках каждой из названных групп в порядке их ввода в БД URUB, а не по алфавиту, как требовал ГОСТ 7.25-2001 (п. 4.11.2.6). Не соблюдался алфавитный порядок экранного представления терминов и при заполнении библиографом поля дескрипторов в АРМ "Каталогизатор", и при работе читателя с "поиском для умников". Это - существенный недостаток, снижающий комфортность и скорость поисковой работы в массиве из нескольких десятков терминов.

Не соблюдался также единый алфавитный порядок построения всего тезауруса при его выводе на экран (и в текстовый файл) из БД URUB в виде лексико-семантического указателя (основная форма представления тезауруса). Отсутствие в структуре рабочего листа БД URUB полей для система-

тического и категориального индексов (разработанные по методическим канонам наши локальные классификации) не позволяло сегментировать массив тезауруса по этим признакам и выводить соответствующие экранные (и текстовые) формы словаря.

В ходе серии встреч с А. И. Бродовским (в конце марта - начале апреля 2012 г.) последовательно решены обозначенные проблемы.

Для корректного экранного представления терминов (и создания текстового файла) использованы rtf-формат (упорядочение текстовых данных) и средства генератора табличных форм, обеспечившие алфавитный порядок терминов внутри словарной статьи и во всём словаре. Для автоматизации заполнения полей систематических и категориальных индексов сделана заготовка для меню-справочника (позже наполненного нами). В структуре рабочего листа URUB реализовано поле 810 для представления источника дефиниции заглавного термина.

Но самое важное - на основе таблицы соответствия полей из эталонной базы тезауруса в ISIS был перенесён в адекватную по структуре и набору функций базу URUB в ИРБИС массив поискового словаря.

Таблица соответствия полей

Метки поля ISIS	**Метки поля URUB**	**Примечание**
020	-	MFN
007	9	Систематический индекс
012	91	Категориальный индекс
001	1	Термин
002	- [перенос термина не требуется - он отображается в ИРБИС на просмотре, поскольку	Синоним-дескриптор (полный) - обратная ссылка

	связь с конкретным термином зафиксирована в самостоят. ссылочной записи]	
006	- [перенос термина не требуется]	Синоним-недескриптор (передается комбинацией дескрипторов)
015	- [перенос термина не требуется]	Синоним-недескриптор (передается одним из альтернативных), обратная ссылка
003	4	Вышестоящий термин
004	100 [перенос термина не требуется]	Нижестоящий дескриптор (удаляется из поля 100 URUB при сохранении)
005	5	Ассоциативные дескрипторы (см. также)
021	-	Идентификатор ("постоянный MFN")
009	3	Примечание

Продолжение таблицы

Метки поля ISIS	**Метки поля URUB**	**Примечание**

009, текст в квадратных скобках [34]	810	Источник информации
010	6	См. (Дескриптор-синоним к недескриптору-заглавному термину) Выводится со словами: **см.**
011	8	исп.А - Альтернативные дескрипторы
008	7	исп.К - Дескрипторы, употребляющиеся комбинацией
022	10	Поле для условного обозначения (да/"пусто") - редактор тезауруса отмечает запись, в которой имеется 2 и более вышестоящих термина разной категории (ШОКОЛАД - КОНДИТЕРСКИЕ ИЗДЕЛИЯ; НАПИТКИ).
072	907^A	gggg-mm-01

В настоящее время идёт тестирование и редактирование эталонного тезауруса в ИРБИС. В ближайшее время мы планируем перенести весь корпус файлов базы URUB на сервер КСОБ. Полная реализация тезауруса в ИРБИС позволяет осуществлять весь технологический цикл ведения библиографической и лексикографической БД в рамках одной программы, законсервировать ведение тезауруса в ISIS (MS DOS), использовать более совершенный механизм создания поискового образа документа и осуществления поиска. А это, в конечном счёте, повышает эффективность использования одного из значимых ресурсов КСОБ.

В заключение отметим очень важный аспект проблемы. Библиотекари-библиографы не всегда обладают необходимыми знаниями в области программного обеспечения автоматизированных систем, с которыми работают. Но они должны учиться правильно ставить задачи перед специалистами.

Выражаем благодарность Александру Иосифовичу Бродовскому за принципиальное решение проблемы реализации тезауруса в системе ИРБИС. Надеемся также, что найденное решение со временем войдет в дистрибутив ИРБИС - одной из наиболее распространённых автоматизированных библиотечных систем.

Список источников

1. **ГОСТ 7.59-2003.** Индексирование документов. Общие требования к систематизации и предметизации // Сборник основных российских стандартов по библиотечно-информационной деятельности. - С.-Петербург.: Профессия, 2006. - С. 261-262.

2. **ГОСТ 7.25-2001.** Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления // Там же. - С. 217.

3. **Прозоров И. Е.** Ведение политематического тезауруса в системе автоматизации библиотек ИРБИС: опыт Центральной городской публичной библиотеки им. В. В. Маяковского / И. Е. Прозоров, Е. П. Здрелюк // Науч. и техн. б-ки. - 2011. - N 11. - С. 114-123.