УДК 026.06

И. Е. Прозоров, Е. П. Здрелюк

Ведение политематического тезауруса в системе автоматизации библиотек ИРБИС: опыт Центральной городской публичной библиотеки им. В. В. Маяковского

На примере ведения тезауруса Корпоративной аналитической библиографической базы данных ЦГПБ им. В. В. Маяковского (С.-Петербург) рассматриваются проблемы, связанные с переходом на систему ИРБИС, сохранением принятой методики индексирования и поиска. Показан вклад специалистов ГПНТБ России в обеспечение функциональных возможностей авторитетной базы тезауруса в ИРБИС.

Ключевые слова: система автоматизации библиотек ИРБИС, Центральная городская публичная библиотека им. В. В. Маяковского, политематические тезаурусы, авторитетные файлы, тематический поиск.

Качество библиографических баз данных и электронных каталогов во многом зависит от программных возможностей АБИС, качества информационно-поискового словаря и созданных на его основе поисковых образов документов (ПОД). Стандартом 7.59–2003 определены общие требования к информационно-поисковым языкам: полнота и точность передачи содержания документа, возможность многоаспектного индексирования, простота и удобство индексирования и поиска [**1**]. В этом отношении дескрипторный ИПЯ представляет собой эффективное средство создания сколь угодно сложного ПОД на основе относительно компактного словаря. Адекватный информационный образ документа формируется из дескрипторов (слов, словосочетаний), как дом из кирпичиков. В то же время, каждый значимый аспект содержания (географический признак, персоналия, событие, организация, предмет и др.) выступает самостоятельной точкой доступа.

Использование тезауруса дает возможность произвольной комбинации терминов как при индексировании, так и при формулировке пользователем поискового запроса. Это принципиально невозможно, например, в сложной предметной рубрике со строго заданной синтагматической связью заголовка с подрубриками. Кроме того, сложная адекватная предметная рубрика стремится передать содержание конкретного документа, что ограничивает возможность ее последующего использования. А это приводит к неоправданному росту поискового словаря.

Простота использования тезауруса может быть сравнима с простотой применения ключевых слов, с той оговоркой, что в тезаурусе, как контролируемом информационно-поисковом словаре, достигнута однозначность терминов путем устранения синонимии, полисемии, омонимии. Кроме того, в тезаурусе поисковый термин предстает в определенном смысловом контексте – в окружении терминов, находящихся с ним в родовидовых, синонимических, ассоциативных отношениях. Иными словами, дескриптор – это не отдельно взятое понятие, а часть наглядно представленной системы знания с возможностью перехода от более общих понятий к более частным и обратно, навигации между терминами одной тематической области, но относящимися к разным лингвистическим категориям (предметам, процессам, явлениям и т.п.).

В тезаурусе реализованы ссылки от непринятых (ненормативных) терминов-синонимов к нормативным терминам или их комбинации либо к перечню возможных альтернативных терминов, которые варьируются в зависимости от смыслового аспекта документа. Таким образом, тезаурус, будучи вербальным поисковым средством, в какой-то мере реализует возможности классификационного ИПЯ по расширению или последовательной локализации искомой предметной области.

Однако реализация тезауруса в автоматизированной системе сталкивается с серьезными ограничениями. Даже ведущие отечественные разработчики тезаурусов (ИНИОН РАН, ЦНСХБ и др.) не располагают всем необходимым набором функций автоматизированной работы с тезаурусами.

Полноценное представление тезауруса предполагает:

- ведение библиографической и словарной (лексикографической) БД в рамках одной АБИС;

- возможность обращения каталогизатора из соответствующего поля библиографической базы к словарной базе для выбора необходимого термина ПОД;
- возможность обращения пользователя к базе тезауруса для формирования поискового запроса при поиске в библиографической базе;
- адекватное представление терминов тезауруса со всеми многообразными связями во фрагменте поискового словаря, видимом каталогизатору при создании ПОД и пользователю при поиске;
- наличие в АБИС полей для реализации тезауруса в соответствии с требованиями ГОСТа.

Для каждого информационно-поискового словаря, в том числе тезауруса, могут устанавливаться свои требования к глубине анализа содержания документов, специфичности используемой терминологии и полноте отражения смысловых связей между терминами. Они определяются задачами ведения базы данных и контингентом ее пользователей.

Библиографическая служба Центральной городской публичной библиотеки им. В. В. Маяковского имеет успешный опыт применения политематического тезауруса для ведения библиографической БД статей на протяжении 17 лет. Неоднократно опыт ведения тезауруса освещался в профессиональной печати и на конференциях [**2**–**4**], на форуме сайта Ассоциации ЭБНИТ (раздел «АРМ Каталогизатор») [**5**].

С 1994 г. по настоящее время эталон тезауруса ведется в оболочке CDS/ISIS/M MSDOS (версия 3.0) с использованием прикладных программ для обработки словарных статей. Изначально библиографическая БД велась посредством этого же программного обеспечения. CDS/ISIS/M позволяет проводить сравнение словарей базы тезауруса и библиографической БД, выявляя новые термины. А при создании новой словарной статьи тезауруса автоматически проставляются обратные ссылки во всех связанных словарных статьях: новый термин включается в соответствующую группу лексических единиц согласно характеру их связи с заглавным термином конкретной статьи. Эта АБИС позволяет легко вносить необходимые изменения в структуру рабочего листа базы тезауруса, представлять на экране тезаурус в виде алфавитного лексико-семантического словаря (основная форма) и выводить его в текстовом формате.

Для ведения тезауруса большое значение имеет возможность сортировки и представления на экране терминов согласно систематическим и категориальным индексам, что позволяет выявлять термины по отраслевому признаку или по лексической категории (предметы, процессы, явления и т.п.).

В 2001 г. ЦГПБ им. В. В. Маяковского начала работать в системе ИРБИС: обеспечение самостоятельного доступа читателей к электронным ресурсам потребовало выбора АБИС с дружественным интерфейсом. Однако возможности ИРБИС на тот момент не отвечали реальным потребностям ведения контролируемых информационно-поисковых словарей, особенно тезауруса. Перед нами встала задача поддержания принятой методики индексирования и поиска. В это время тезаурус по-прежнему вели в CDS/ISIS/M, а библиографы при индексировании использовали его текстовую версию. Для отражения имен собственных (дескрипторов последнего уровня) использовали поля предметных рубрик: тематической, географической, имен лиц, коллективов. Эта группа терминов в рамках тезауруса изначально велась в отдельной лексикографической базе и выводилась в самостоятельный словарь. Для терминов основного словаря в библиографической БД первоначально использовалось поле 610 ненормированных ключевых слов, а позже – появившееся поле 965 «Дескрипторы».

Существенным недостатком ИРБИС в те годы было то, что доступный пользователям словарный поиск низводил дескрипторы до уровня ключевых слов. Развитие средств ведения лексикографических БД и поиска через авторитетный файл («Поиска для умников») стало, на наш взгляд, самым существенным достижением системы ИРБИС.

Важным этапом работы библиографической службы нашей библиотеки в системе ИРБИС стал 2006 г. Из словаря имен собственных персоналии (имя лица) были перенесены в авторитетный файл имен в ИРБИС. Механизм функционирования авторитетных файлов предметных рубрик хорошо себя зарекомендовал, вскоре в ИРБИС были переведены географические рубрики.

Но самое главное – тогда же, в 2006 г., основной массив тезауруса (более 7 тыс. терминов) был перенесен в авторитетную лексикографическую БД «Тезаурус» в ИРБИС. С того времени при работе в библиографической БД библиографы, создавая поисковый образ документа, могли одномоментно выбрать весь необходимый перечень дескрипторов из тезауруса, открывающегося в поле 965. Это существенно

ускорило темпы аналитической обработки изданий, сократив количество ошибок ввода. И у библиографов, и у читателей через «Поиск для умников» появилась возможность увидеть дерево тезауруса и место отдельного термина в нем, а также ознакомиться с примечаниями, синонимами, рекомендуемой комбинацией дескрипторов и т.д. непосредственно в системе ИРБИС.

Однако целый ряд функциональных ограничений не позволил перенести процесс ведения тезауруса в ИРБИС. Формат базы тезауруса в ИРБИС не обеспечивал отражения всех необходимых семантических связей. В сложных случаях у библиографов возникала необходимость навести справки по регулярно обновляемой текстовой версии тезауруса в MS Word.

Для обновления тезауруса в ИРБИС из эталонной базы в CDS/ISIS/M, в текстовом формате, извлекались новые авторитетные записи и те, в которые были введены изменения. Затем вся информация последовательно переносилась в авторитетную базу системы ИРБИС.

При создании новых словарных статей заполнялись поля: Основной термин; Вид – дескриптор/недескриптор; Вышестоящий дескриптор (вводится из справочника); Ассоциативные дескрипторы (вводится из справочника); Примечание русское; Уровень дескриптора в иерархии. А для аскрипторов заполнялись поля: для недескрипторов – синонимическая связь (вводится из справочника); для недескрипторов – комбинация дескрипторов (вводится из справочника); для недескрипторов – перечень дескрипторов (вводится из справочника).

Изменения в дескрипторных статьях как правило затрагивали вышестоящий дескриптор, поле ассоциативных дескрипторов и примечание. К сожалению, на этом этапе работы вносить изменения в перечень нижестоящих дескрипторов не представлялось возможным из-за отсутствия соответствующего поля.

Для полноценной реализации тезауруса в системе ИРБИС требовалось:

1. Обеспечить в формате базы «Тезаурус» повторяемость поля для вышестоящих дескрипторов: достаточно часто у дескриптора необходимо отразить два–три вышестоящих термина. Например, в нашем тезаурусе к дескриптору абстракционизм имеется три вышестоящих – авангардизм, модернизм, художественные течения и стили; с дескриптором земельные участки связаны три вышестоящих – недвижимость, основные средства, средства производства; с дескриптором национальный доход связаны два термина – *доходы* и *экономические показатели*;

2. Создать повторяющееся поле для нижестоящих дескрипторов. Нижестоящие термины не отражались в окне «Запись полностью» при работе с вложенным окном поля 965 «Дескрипторы», что не позволяло библиографу проводить полноценный поиск и отбор терминов непосредственно при заполнении этого поля. Доступный пользователю навигатор «Поиска для умников» отражал нижестоящие термины к дескриптору (поскольку в тезаурусе каждый нижестоящий термин в своей дескрипторной статье содержал конкретный вышестоящий термин), но выводил их в произвольной последовательности, а не по алфавиту;

3. Создать повторяющееся поле для синонимов-аскрипторов в составе листа ввода данных, функционально соответствовавших 410 полю формата авторитетных файлов в ИРБИС;

4. Обеспечить автоматическую корректировку всех связанных словарных статей при редактировании одной из записей базы тезауруса или при создании новой (простановка обратных связей).

Наконец, важно было обеспечить адекватное указанным лексическим связям отображение словаря на экране – то, что позволяло бы использовать возможности, заложенные в структуру базы тезауруса.

Среди предложений по модернизации модуля «Тезаурус» в АБИС (на примере ИРБИС), высказанных нами на Международном библиографическом конгрессе в Санкт-Петербурге в сентябре 2010 г., были и такие:

- создание в структуре базы тезауруса фиксированного поля для идентификационного номера дескриптора (независимого от изменчивых MFN);
- реализация возможности автоматической коррекции поисковых образов в библиографической БД в зависимости от изменений в базе тезауруса;
- создание полей для категориальных и систематических индексов.

Сложность ситуации обусловила необходимость выбора приоритетных, наиболее актуальных проблем. Огромную поддержку при их решении оказал заведующий отделом разработки и совершенствования АБИС ГПНТБ России Александр Иосифович Бродовский. В ходе встреч, состоявшихся в ГПНТБ России и ЦГПБ им. В. В. Маяковского в декабре 2010 – январе 2011 гг., он предложил ряд конкретных решений.

Прежде всего, была отмечена невозможность отображения на экране всех заявленных связей между терминами средствами базы «Тезаурус». Реализация требования повторяемости поля для множества вышестоящих терминов требовала бы переработки идеологии программного обеспечения функций этой БД. Поэтому для ведения тезауруса предложено использовать предназначенный для создания различных лексикографических БД универсальный рубрикатор URUB. Были внесены изменения в служебные файлы, содержащие параметрические описания БД URUB. Поля базы были настроены на представление тезауруса и переименованы в соответствии с функциональным назначением: например, поле «Ключевые слова» переименовано в «Примечание» (для дефиниций), добавлены повторяющиеся поля для вышестоящих и нижестоящих терминов.

Для корректного отображения словарных статей тезауруса в окне полного описания АРМ «Каталогизатор» в базе URUB, а также фрагментов словаря при работе с тезаурусом во вложенном окне поля 965 «Дескрипторы» библиографической БД были использованы средства ИРБИС-навигатора на основе HTML, как в «Поиске для умников». Именно средства гипертекстовой разметки позволили буквально «нарисовать» необходимую структуру тезауруса, обеспечив взаимный переход по терминам-гиперссылкам.

Стандартом 7.25-2001 «Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления» [**6**] определены требования к его организации, среди которых:

- взаимность всех ссылок тезауруса (п. 4.5.7);
- в статье нижестоящего дескриптора должна быть ссылка на вышестоящий термин, а в статье вышестоящего – на нижестоящий;
- разделение синонимов-недескрипторов (аскрипторов) на три группы: 1) недескрипторы, эквивалентные одному принятому дескриптору (характер связи в словаре отображается ссылкой «Смотри»); 2) недескрипторы, значение которых передается комбинацией дескрипторов (характер связи отражается ссылкой «Смотри комбинацию»); 3) недескрипторы, заменяемые одним дескриптором из нескольких возможных альтернативных (характер связи отражается ссылкой «Смотри альтернативу»);
- отражение в дескрипторных статьях синонимов-недескрипторов с указанием характера связи;
- алфавитная последовательность словарных статей в лексико-семантическом указателе (основной вид представления тезауруса);
- указание характера связи с заглавным термином статьи при представлении терминов в словарных статьях;
- соблюдение следующей последовательности терминов в дескрипторной статье: 1) заглавный дескриптор; 2) лексическое примечание (дефиниция); 3) синонимы-недескрипторы; 4) вышестоящие дескрипторы; 5) нижестоящие дескрипторы; 6) ассоциативные дескрипторы;
- алфавитный порядок группировки внутри каждой лексической группы;
- наконец, принципиально важным требованием является представление тезауруса на видеотерминалах в соответствии с основными требованиями оформления по данному ГОСТу (п. 4.11.3).

В результате встреч с А. И. Бродовским массив тезауруса был инсталлирован в базу URUB, данные распределены по полям (названия и функциональные значения полей были предложены специалистами библиографической службы ЦГПБ им. В. В. Маяковского):

поле 1: Термин – для заглавного термина словарной статьи;

поле 3: Примечание – для дефиниции термина;

поле 4: Вышестоящий дескриптор – повторяющееся поле;

поле 5: Ассоциативные дескрипторы (ссылки «Смотри также») – повторяющееся поле;

поле 6: Для недескрипторов – Синоним (См.) – для указания дескриптора, передающего смысл заглавного

термина-недескриптора;

поле 7: Для недескрипторов – Комбинация терминов (См.) – повторяющееся поле для дескрипторов, совместно передающих смысл заглавного термина-недескриптора;

поле 8: Для недескрипторов – Перечень альтернативных дескрипторов (См.) – повторяющееся поле для нескольких дескрипторов, предложенных на выбор для передачи смысла недескриптора;

поле 100: Нижестоящие (удаляются после сохранения) – поле для отображения нижестоящих терминов при создании дескрипторной статьи. Внесенные термины отображаются в окне «Полное описание» до сохранения записи как неактивные ссылки для целостного отображения дескрипторной статьи, после сохранения записи вышестоящий дескриптор отображается в соответствующем поле каждого указанного здесь дескриптора как вышестоящий, нижестоящие термины в создаваемой дескрипторной статье отображаются на экране как активные ссылки;

поле 907: Дата.

По нашей просьбе реализована автоматическая двусторонняя связь между ассоциативными терминами так, чтобы при создании новой дескрипторной статьи (или редактировании существующей) в соответствующие дескрипторные статьи на правах ассоциативного термина включался бы новый (или редактируемый) дескриптор.

Поля 6–8 для отражения ссылок заполняются только в аскрипторных, т.е. ссылочных статьях, где заглавный термин является недескриптором (сам факт заполнения одного из этих полей определяет соответствующий статус заглавного термина).

Первоначально реализованный в URUB тезаурус не дифференцировал виды ссылок, сопровождая их единым комментарием «Смотри», что было существенным недостатком и создавало угрозу появления ошибок при индексировании. Так, например, термин *порядок исчисления заработка* выражается в тезаурусе термином *оплата труда*. Термин *полиэтилен* передается, в зависимости от содержания документа, дескрипторами: *полимеры*, или *потребительские товары*, или *химические товары*. А, например, термин *почтовые переводы* передается в поисковом образе совместным использованием дескипторов *переводы [фин.]* и *почта*. Между тем по ГОСТу 7.25-2001, это должны быть разные виды ссылок.

Другой временной проблемой было то, что при первоначальной инсталляции тезауруса были потеряны дефиниции, что также создавало угрозу точности отбора терминов при формировании поисковых образов в библиографической базе.

В ходе встречи с А. И. Бродовским были решены проблемы дифференцированного отображения ссылок и представления в тезаурусе лексического примечания. Приведен в соответствие с требованиями ГОСТа формат вывода на экран словарной статьи: в окне «Полного описания» в базе URUB, во вложенном окне поля 965 «Дескрипторы» библиографической базы, в окне «Поиска для умников».

Решена также проблема алфавитной группировки терминов внутри каждой группы лексических единиц словарной статьи, что создало условия нормальной работы с тезаурусом. Это существенное требование, особенно когда число терминов может достигать нескольких десятков (ассоциативных и нижестоящих дескрипторов).

Отметим и такое, очень существенное для качества создаваемой библиографической БД, приобретение, как исключение возможности ввода в поисковый образ документа недескрипторов и вспомогательных терминов, снабженных систематическими индексами и служащих для образования дерева терминов. (Эти термины выполнили вспомогательную роль при переводе тезауруса из CDS/ISIS/M в ИРБИС, во многих случаях дескрипторы были присоединены к той или иной отраслевой группе достаточно условно.)

Важно отметить, что все нововведения сразу же были введены в производственный процесс библиографических служб публичных библиотек Санкт-Петербурга, объединенных в единую корпоративную сеть.

Итог: выбор тезауруса как эффективного средства индексирования и тематического поиска, создание и

поддержание информационно-поискового словаря, отвечающего требованиям ГОСТа, в свое время создали важные предпосылки обеспечения качества библиографической БД. Реализация словаря адекватными программными средствами в настоящее время позволяет создавать эффективно работающий механизм анализа документов, индексирования и поиска, соответствующий интересам пользователей публичных библиотек мегаполиса.

В настоящее время осуществляется дальнейшая доработка авторитетной базы URUB, планируется перенос процесса ведения тезауруса в ИРБИС.

С разрешения А. И. Бродовского мы готовы предоставить служебные файлы базы URUB всем заинтересованным специалистам (ibo@pl.spb.ru; prozorov@pl.spb.ru).

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. 262.

2. **Оранская Л. И.** Некоторые особенности использования дескрипторного поискового языка в библиографической ИПС универсальной библиотеки // Науч. и техн. б-ки. – 1997. – № 9. – С. 13–23; 1998. – № 1. – С. 153.

3. **Сухарева М. Н.** Проблемы отражения краеведческих материалов в библиографической базе данных // Проблемы краеведческой деятельности библиотек / РНБ, Новгород. обл. универс. науч. б-ка. – С.-Петербург, 2003. – С. 107–111.

4. **Прозоров И. Е.** Координатное индексирование и возможности координатного поиска в библиотечных информационно-поисковых системах / Прозоров И. Е., Кисарова М. Е., Сухарева М. Н. // Лингвистическое обеспечение информационных ресурсов библиотек, музеев, архивов и других учреждений культуры / ЦГПБ им. В. В. Маяковского. – С.-Петербург, 2008. – С. 95–111.

5. **Тезаурус** в ИРБИС: проблемы применения / [Прозоров] Иван Евгеньевич // Ассоциация ЭБНИТ. Форумы. ИРБИС. АРМ-Каталогизатор. – Москва, 2010–2011. – Режим доступа: http://irbis.gpntb.ru/read.php?10,46024

6. **ГОСТ 7.25–2001.** Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления // Сборник основных российских стандартов по библиотечно-информационной деятельности. – С.-Петербург : Профессия, 2006. – С. 201–219.