Галина Анатольевна Скарук
кандидат педагогических наук, старший научный сотрудник
Государственной публичной научно-технической библиотеки Сибирского отделения Российской академии наук

Проблемы создания и поддержания лексикографических баз данных для электронного каталога

The problems of creating and maintaining a lexicographical database for the electronic catalog

Calina A. Skaruk, senior research worker, cand. of pedagog. sciences,
SPSTL SB RAS

Рассматриваются причины недостаточного использования лексикографических баз данных в электронных каталогах. Анализируются проблемы создания и ведения двух баз данных , создаваемых в ГПНТБ СО РАН,: авторитетного файла предметных рубрик и алфавитно-предметного указателя к каталогам. Представляются пути решения этих проблем.
Ключевые слова: лексикографические базы данных, словари информационно-поисковых языков, предметные рубрики, авторитетный файл, Библиотечно-библиографическая классификация, электронный алфавитно-предметный указатель.

The reasons of underutilization of lexicographical database in electronic catalogs are considered. The problems of creating and maintaining two databases created in SPSTL SB RAS: authority file of subject headings and Alphabetical Index for the catalogs. Presents solutions to these problems.
Keywords: lexicographical databases, dictionaries of information retrieval languages, subject headings, authority file, bibliographic classification, electronic Alphabetical Index.

Электронные каталоги (ЭК) по-прежнему остаются основными библиографическими ресурсами библиотек, отражающими содержание их фондов. С развитием ЭК развиваются и лексикографические базы в их составе, все большее их количество необходимо для успешного индексирования и поиска документов. Расширяются их функции.

Напомним, что под лексикографическими БД мы имеем ввиду базы данных (БД), представляющие собой различные машиночитаемые словарные массивы, объектом описания в которых является лексическая единица. Это могут быть тезаурусы, рубрикаторы, терминологические словари, таблицы классификации т. д. Иначе говоря – в основном это машиночитаемые словари ИПЯ,

Практически все системы автоматизации библиотек (САБ) располагают возможностями применения нормативных словарей информационно-поисковых языков и формирования поисковых предписаний на их основе. Почти все САБ декларируют в описаниях возможность создания собственных или подключения внешних авторитетных файлов разных типов.

Но на практике эти возможности почти не реализованы. Встроенные системы для работы с классификационными ИПЯ и предметными рубриками есть только в библиотеках, использующих ИРБИС или «Руслан». Проблемы реализации названных возможностей в конкретных библиотеках связаны не только с недостатком финансовых и кадровых ресурсов.

Нам кажется, что основная причина здесь – отсутствие лингвистических средств для наполнения баз.

Серьезнейшая проблема лингвистического обеспечение электронных каталогов в России − отсутствие единых общепризнанных конкретных словарей ИПЯ, использование которых обеспечит совместимость электронных каталогов и баз данных, генерируемых библиотеками нашей страны. Полные таблицы Библиотечно-библиографической классификации, на основе которых построены каталоги трех крупнейших библиотек России,

не будут больше издаваться. Единый Авторитетный файл РНБ создан на основе ретроконверсии карточного предметного каталога библиотеки и, судя по информации на сайте, планового редактирования его в ближайшее время не предвидится, также. как и его печатного издания. Давно затихли разговоры о Российском национальном авторитетном файле. Поэтому многие библиотеки, решая проблемы поисковой лингвистики ЭК, идут своими путями.

Наиболее простой и распространенный из них – ключевые слова. Основной аргумент в их пользу – экономичность технологии. Нет необходимости создавать специальные лексикографические БД и специальные инструменты поиска. На самом же деле это – ситуативное решение, которое в конечном итоге приведет в тупик.

Таким образом, существует проблема отсутствия исходного материала для наполнения лексикографических БД. Часто пользователям для работы предоставляются устаревшие списки предметных рубрик, классификационные таблицы.

Учитывая ряд обстоятельств, которые описаны в статьях [1, 2], ГПНТБ СО РАН приняла решение создавать собственный авторитетный файл в ИРБИС64, на основе существующей в системе БД «Предметные заголовки», поэтому на практике получила возможность проверить, как работает авторитетный файл (АФ) предметных рубрик в этой системе.

База “Предметные заголовки” представляет собой традиционную базу ИРБИС со специально сформированной структурой.

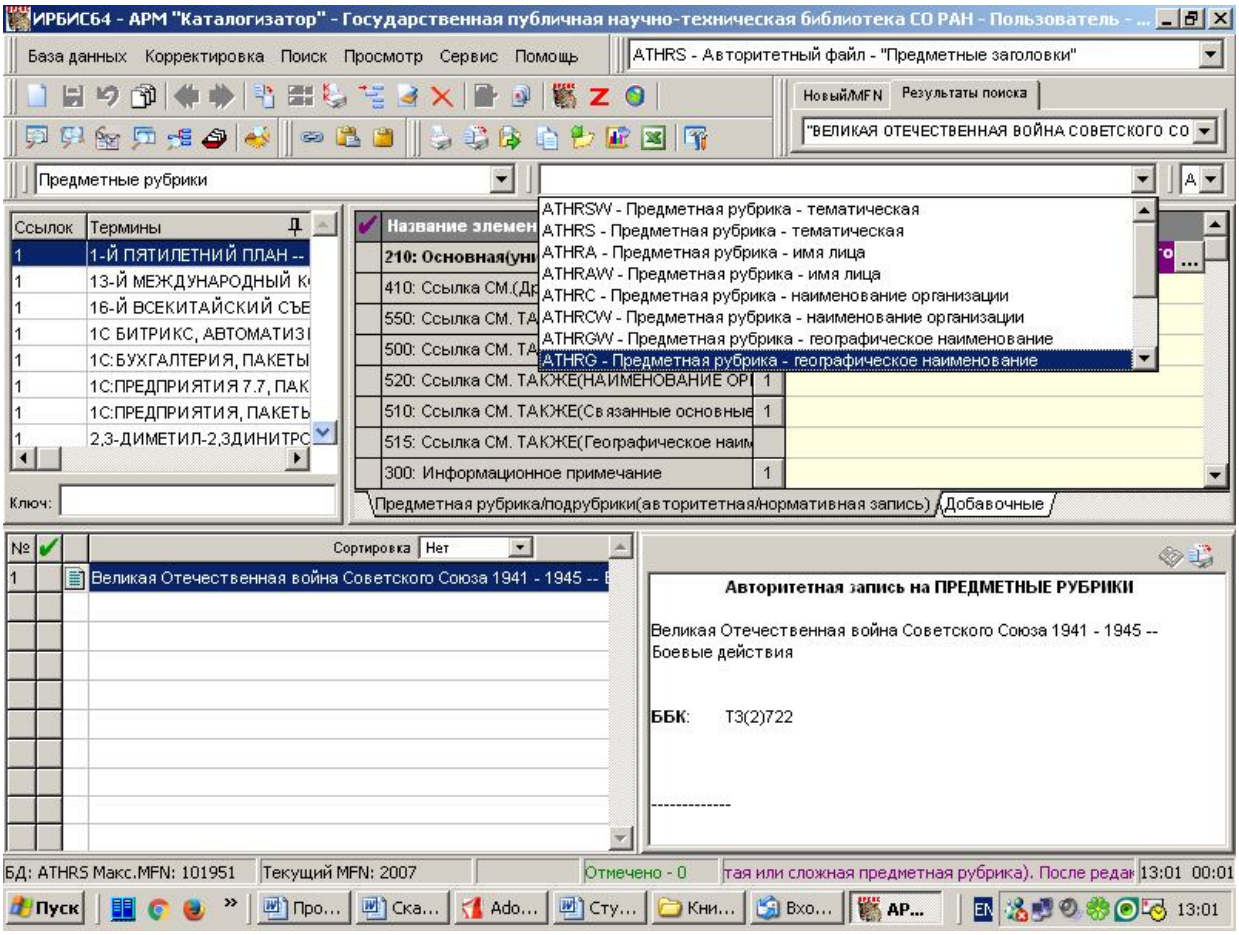

Рис.1. База данных «Предметные заголовки» в ИРБИС64.

Система поиска такая же, как и в других базах. Поиск возможен по ключевым словам, полному тексту предметных рубрик, тематическим, географическим, хронологическим, формальным подзаголовкам, по соответствующему рубрике индексу ББК, шифру предметной рубрики в АФ. Особо выделяются ПР, нуждающиеся в редактировании. В ИРБИС64 разработана структурная основа АФ предметных рубрик, сформированы основные поля и подполя. Но при практической реализации базы возникло большое количество затруднений.

1. Отсутствие в структуре формата ввода некоторых подполей для принятой формы ПР.

Методика предметизации РНБ [3] предлагает правила разбиения словосочетаний на заголовки, а также порядок следования заголовка и подзаголовков предметных рубрик, несколько отличающиеся от заданных в

формате ввода ИРБИС64. Например. Дата после исторического события в RUSMARCe должна фигурировать как отдельный подзаголовок, у нас подполя для этого не предусмотрено.

Великая Отечественная война − 1941 - 1945 − Военные операции.
Авторитетная запись рубрики в формате RUSMARC в ЕАФ РНБ:

250 ##$aВеликая Отечественная война$z1941 - 1945$xВоенные операции

Великая Отечественная война Советского Союза 1941 - 1945 − Боевые действия.

Авторитетная запись рубрики в оптимизированном формате ИРБИС64:

^aВеликая Отечественная война Советского Союза 1941 – 1945^bБоевые действия

Еще один пример:

210 01$aРоссийская Федерация$bПрезидент$bАдминистрация$bБиблиотека

Таких форм ввода с повторяющимися полями в ИРБИСе вообще нет.

2. Несовпадение структуры подполей в рабочих листах для блока анализа содержания документа (--6) библиографического формата и для блока принятых точек доступа (--2) формата RUSMARC-Authorities. В связи с этим нам пришлось дорабатывать рабочие листы для ввода имен лиц и наименований организаций.

3. Недостаточная наглядность форматов вывода авторитетных данных создаст определенные неудобства при использовании АФ читателями.

5. Недостаточная разработанность блока ссылок и примечаний.

Это выразилось в отсутствии подполей для ряда элементов ссылок, неправильном с точки зрения принципов RUSMARCa формате вывода и просмотра. Эти позиции также пришлось дорабатывать.

6. Недостаточно наглядное представление ссылок.

Ссылочные записи в ИРБИ64 выглядят непривычным для пользователя каталогов образом и выводятся при индексировании только по требованию, поэтому работать с ними неудобно. Эта часть программы оставлена пока без изменений.

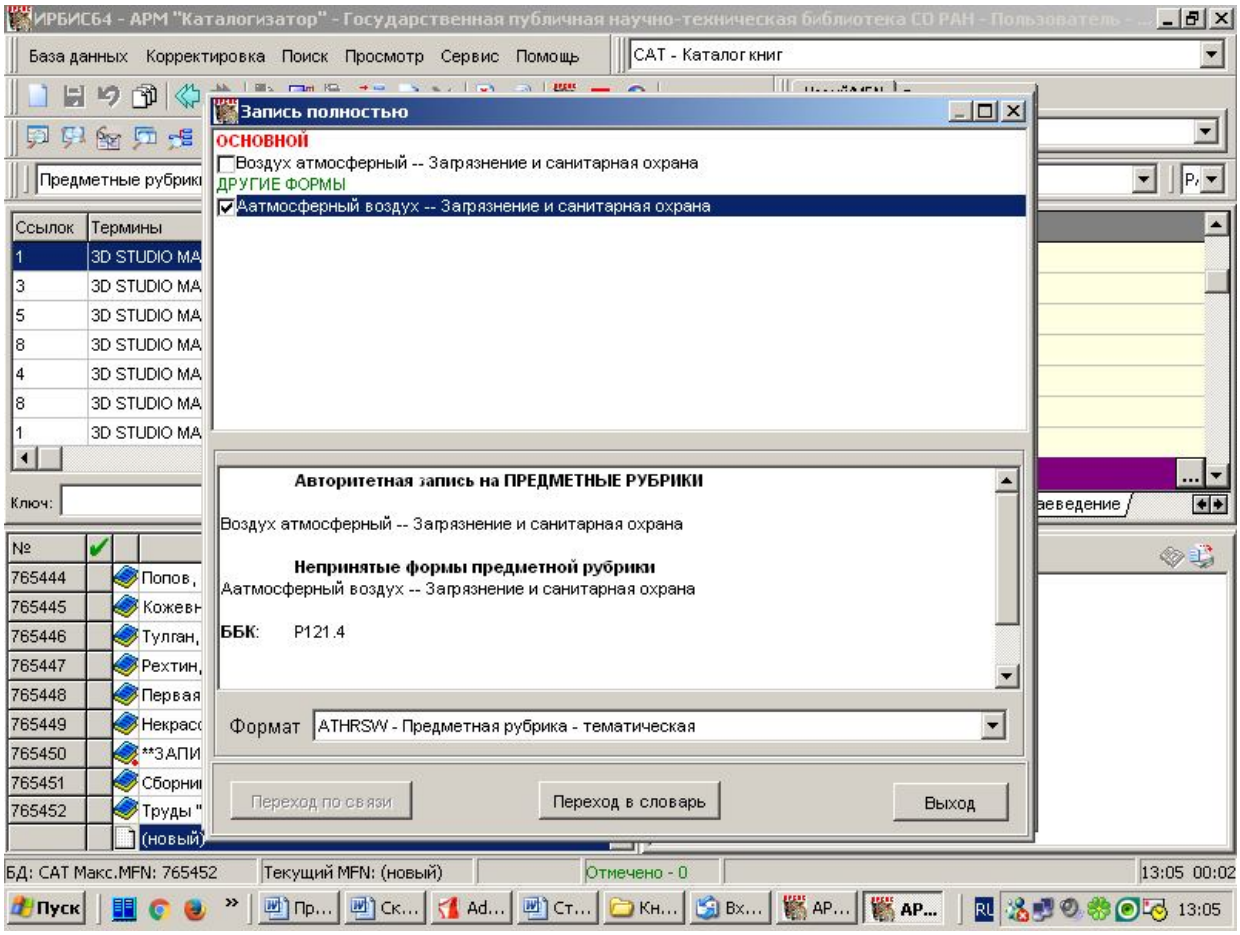

Рис. 2. Представление ссылки «см.» в Авторитетном файле предметных рубрик в ИРБИС64.

7. Большого количества времени потребовала загрузка в авторитетный файл рубрик из библиографического файла электронного каталога.

Сегодня блок файла уже работает в режиме текущей эксплуатации.

В этой области основные затруднения связаны с организацией управления АФ, прежде всего с его пополнением за счет вновь введенных в ЭК ПР. Предполагалось, что такие рубрики будут сразу после введения каждым предметизатором попадать в авторитетный файл, регистрироваться как нуждающиеся в редактировании и, после снятия редактором пометки о необходимости просмотра, использоваться всеми индексаторами. Разработана программа реализации этой процедуры, но проблема до сих пор так и не решена.

Важным достижением мы считаем создание системы редактирования ПР в библиографических записях непосредственно через АФ. Предметные рубрики, исправленные в авторитетных записях, автоматически исправляются во всех БЗ электронного каталога.

Серьезным подспорьем в предметизации служит возможность поиска предметных рубрик по связанным с ними индексам ББК. ББК в этом случае служит инструментом структурирования массива рубрик. Поиск же в неструктурированном массиве АЗ большого объема требует больших временных и интеллектуальных затрат.

Помимо чисто технических трудностей, ряд проблем был связан с методикой предметизации. Перед копированием в авторитетный файл предметные рубрики в библиографическом файле электронного каталога подлежали редактированию. Этот процесс оказался гораздо сложнее и длительнее, чем мы предполагали. В 1990-е годы четкой и конкретной методики предметизации не существовало. В ЭК, ведущемся ГПНТБ СО РАН с 1991 г., рубрики приводились на основе рубрик АПУ к систематическому каталогу, но по разным методикам, в разных системах.

Серьезной задачей стала выработка методики предметизации на основе методики Российской национальной библиотеки. Ее внедрение потребовало довольно длительной планомерной работы.

В поддержку этой работы была создана БД «Электронная картотека методических решений по предметизации». Она ведется на постоянной основе, методические решения обсуждаются на методическом совете секторов систематизации и вводятся в базу. Для нее была разработана специальная система поиска.

Еще одна лексикографическая база данных, создание которой началось в прошлом году, − «Алфавитно-предметный указатель к систематическому каталогу». К этому нас подвигла необходимость организации поиска по индексам БКК в базе на основе имиджей карточек систематического каталога и электронном каталоге по тем темам, которые не войдут в АФ ПР (сейчас не

изучаемым, «устаревшим», об областях деятельности, которыми в настоящее время не занимаются и пр.).

База формируется на основе распознавания образов карточек карточного АПУ, разнесения по подполям и редактирования. Эта БД функционирует как обычная БД ИРБИСа, только вместо библиографической информации каждая запись содержит рубрику АПУ и соответствующий ей индекс ББК.

Основные проблемы здесь – выработка и реализация механизмов, с помощью которых были распознаны и разнесены по полям заголовки и подзаголовки рубрик, методические указания. Кроме того, предстоит большой объем «ручной» работы по редактированию ошибок распознавания.

База АПУ по сравнению с карточным АПУ будет обладать многими преимуществами. Будет возможен поиск не только по первому слову рубрики, а по любому слову, а также поиск рубрик по индексам ББК. Это позволит объединить в ней непосредственно сам АПУ и систематические указатели к нему.

При решении этих задач нас, как и другие библиотеки, затронула проблема дефицита кадров. Объемные задачи решаются малыми силами. Тем не менее, необходимость лексикографических баз данных отчетливо осознается и предпринимаются меры по оптимизации усилий.

Литература:

1. Скарук Г. А. Перспективы создания авторитетного файла предметных рубрик в ГПНТБ СО РАН // Труды ГПНТБ СО РАН. - Новосибирск, 2011. – Вып. 1 : Развитие электронной информационно-библиотечной среды. – С. 295–300.

2. Скарук Г. А. Авторитетный файл предметных рубрик: новые возможности индексирования и поиска // Научные библиотеки: вчера, сегодня, завтра. – Новосибирск, 2013. – С. 269–277. – (Труды ГПНТБ СО РАН ; вып. 4).

3. Руководство по методике предметизации. Опыт Российской национальной библиотеки / Рос. нац. б-ка, Нац. информ.-библ. центр "ЛИБНЕТ" ; [авт.-сост. : Ю. Г. Селиванова и др.]. – Москва : Фаир пресс [и др.], 2005. - 407 с.