Проблемы библиотечной автоматизации приобретают особую актуальность в условиях технического перевооружения библиотек, их компьютеризации. На наших глазах повсеместная автоматизация библиотечно-библиографических процессов, создание электронных каталогов и баз данных становятся реальностью для большинства учреждений отрасли. Этот процесс отвечает потребностям пользователей получать информацию оперативно и максимально полно независимо от места нахождения библиотеки, ее фондов и подготовленности пользователя.

ИПЯ - ЯЗЫК, КОТОРЫЙ НАДО ЗНАТЬ

О РОЛИ ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ В РАЗВИТИИ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ БИБЛИОТЕК

Лидия ПИРУМОВА,
заместитель директора Центральной научной сельскохозяйственной библиотеки РАСХН, кандидат педагогических наук

Создать условия, при которых читатель может получить доступ к информационно-поисковым системам (ИПС) библиотек различной удаленности и вести эффективный поиск в них, помогают лингвистическое обеспечение (ЛО) и его основная составляющая — информационно-поисковые языки (ИПЯ). Задача библиотек состоит не только в том, чтобы собрать в своих фондах возможно полно документы, но сделать их доступными для пользователя, дать информацию о них и раскрыть информацию, содержащуюся в них. Всему этому способствуют каталоги, базы данных (БД), библиографические и реферативные издания. Информация в них должна быть систематизирована и представлена в таком виде, который позволяет осуществлять быстрый поиск в данных ИПС, БД, электронном каталоге (ЭК).

Любая ИПК включает следующие элементы: информационный массив; ИПЯ, на которой переводится входная информация и запросы; правила этого перевода (индексирование); критерии выдачи, то есть правила сравнения перевода запроса на ИПЯ с результатами перевода на ИПЯ входной информации, определяющие отбор информации, подлежащей выдаче на запрос.

Понятие ЛО шире понятия информационно-поискового языка, поскольку включает их в себя. Лингвистическое обеспечение автоматизированных систем включает ИПЯ, методики индексирования документов и запросов на них, инструкции и методики их ведения и использования, а также средства поддержания ИПЯ в автоматизированной системе.

Средством свертывания информации и смысловой обработки документов является информационно-поисковый язык (ИПЯ) — формализованный искусственный язык, предназначенный для индексирования документов, информационных запросов. Искусственный язык, специально разработанный для автоматизированного поиска, лишен недостатков естественного языка (многозначность, избыточность) и лучше приспособлен для информационного по-

«СРЕДСТВОМ СВЕРТЫВАНИЯ ИНФОРМАЦИИ И СМЫСЛОВОЙ ОБРАБОТКИ ДОКУМЕНТОВ ЯВЛЯЕТСЯ ИНФОРМАЦИОННО-ПОИСКОВЫЙ ЯЗЫК (ИПЯ) — ФОРМАЛИЗОВАННЫЙ ИСКУССТВЕННЫЙ ЯЗЫК, ПРЕДНАЗНАЧЕННЫЙ ДЛЯ ИНДЕКСИРОВАНИЯ ДОКУМЕНТОВ, ИНФОРМАЦИОННЫХ ЗАПРОСОВ».

иска, увеличивая полноту и точность выдачи информации. При создании ИПЯ учитываются требования, которые отвечают его задаче — полноте и точности поиска: однозначность — каждая запись на ИПЯ должна иметь только один смысл, то есть искусственный ИПЯ должен устранять такие недостатки, с точки зрения поиска естественного языка, как полисемия и омонимия; явное выражение полезных для поиска семантических (смысловых) отношений между словами (логических отношений и психологических ассоциаций) ИПЯ; возможность корректировки и дополнения ИПЯ; удобство пользования, ИПЯ должен обладать компактностью записей, способствующих его запоминанию; способность точно идентифицировать предмет, отличить его особенности и описать его с необходимой степенью детализации и глубины.

Семантическое богатство ИПЯ зависит от его терминологической наполненности, структуры построения и от взаимоотношений лексических единиц, составляющих лексику, словарный состав ИПЯ. Лексическая единица (ЛЕ) информационно-поискового языка — это обозначение отдельного понятия, принятое в нем. ЛЕ каждого ИПЯ называются по-разному: в классифицированных системах — это индексы, в языке предметных рубрик — рубрики, в дескрипторных языках — дескрипторы, в языке ключевых слов — ключевое слово. По тому, какие ЛЕ используются в ИПЯ, различают словарные и кодированные ИПЯ. В словарных ИПЯ (тезаурус) используются элементы естественного языка, и перевод на естественный язык не требуется. В кодированных ИПЯ (УДК, ББК) индексы или рубрики сопровождаются таблицей соответствия, то есть каждой ЛЕ на искусственном языке дается словесное ее выражение на естественном языке. Основу лексики любого ИПЯ составляют термины, являющиеся носителями научной информации в текстах документов. Любой ИПЯ создается на основе терминологии определенной области знаний.

Разработка ИПЯ проходит несколько этапов: отбор лексических единиц; процесс нормализации лексики; систематизация и группировка лексики; построение классификационных схем; оформление лексики ИПЯ.

Этап отбора лексических единиц особенно важен в процессе создания информационно-поискового языка, поскольку от него зависят возможности данного ИПЯ: терминологическая наполненность, соответствие уровню развития науки, отражаемой в нем, а значит, и поисковые возможности данного ИПЯ. Отбор ЛЕ происходит в процессе аналитико-синтетической обработки документов на этапе аннотирования, систематизации индексирования.

ИПЯ неразрывно связан с процессом аналитико-синтетической обработки информации, поскольку на этом этапе раскрывается тематическое содержание документа, происходят свертывание информации, представленной в нем, и ее перевод на формализованный язык, позволяющий внести информацию в ЭК, а затем вести в нем поиск. Прежде чем информация предстанет в виде элементов ИПЯ, она проходит семантическую, то есть смысловую обработку. Текст, представленный на естественном языке, анализируется с точки зрения его содержания. В ходе осмысления содержания текста документа человеком (семантической обработки) происходит отбор наиболее значимых, основных тем документа, а затем их перевод с естественного на искусственный язык. При этом точность и полнота перевода зависят от возможностей ИПЯ, от уровня разработки его лексического и терминологического аппарата, наличия правил этого перевода.

Таким образом, именно ИПЯ является основным компонентом любой ИПС, без которой она превращается только в беспорядочный «сундук» информации.

В традиционной ИПС использовались ИПЯ, разработанные для карточных каталогов; наибольшее распространение получили Универсальная десятичная классификация (УДК) и Библиотечно-библиографическая классификация (ББК). Однако использование их в автоматизированных системах пока не обеспечивает эффективного поиска. Вместе с тем существуют ИПЯ, специально разработанные для автоматизированных ИПС и для автоматизированного поиска: рубрикаторы, тезаурусы. При создании электронных каталогов, автоматизированных ИПС перед библиотеками встает задача выбора ЛО и ИПЯ, которые будут использоваться в них.

Как правило, в одной информационно-поисковой системе используются несколько ИПЯ, поэтому встает вопрос об их совместимости. В условиях одной ИПС эта проблема решается, если все документы, входящие в ее документный поток, индексируются на всех ИПЯ, используемых в данной поисковой системе. Для достижения совместимости в одной ИПС следует обеспечить единую методику индексирования на всех ИПЯ этой системы, а также добиться унификации и стандартизации языковых средств и поддерживающих компонентов ЛО.

Использование нескольких ИПЯ в одной ИПС объясняется тем, что каждый из языков предназначен для выполнения определенных функций в ней, а также осознанием того, что не может быть создан единый ИПЯ, выполняющий одновременно все функции лингвистических средств и все задачи, стоящие перед информационно-поисковой системой. Одновременное использование нескольких информационно-поисковых языков обеспечивает быстрый и разнообразный доступ потребителя к информационным ресурсам в зависимости от его знания какого-либо из ИПЯ и от того, какого рода информация ему нужна и для каких целей. Все это относится к решению проблемы узкой совместимости в рамках одной ИПС.

Проблема совместимости средств ЛО различных ИПС стала особенно актуальна с развитием информационных сетей. Поскольку каждая ИПС использует свои ИПЯ, то обмен информацией между информационно-поисковыми системами затруднен из-за несовместимости этих ИПЯ. Различают средства и методы достижения лингвистической совместимости. К средствам ее обеспечения относятся рубрикаторы, классификаторы, библиотечные форматы записи, тезаурусы и нормативные словари, конверторы, необходимые для перевода информации из одной формы ее предоставления в другую. К основным методам совместимости лингвистических средств относят: методологическую совместимость; стандартизацию и унификацию языковых средств; создание общесетевых универсальных ИПЯ; сопряжение языковых средств; методы конверсии языковых средств; сосуществование разных ИПЯ в сети.

Методическая совместимость — это разработка единых принципов

создания и ведения ЛО отдельных ИПС, входящих в одну информационную сеть; разработка нормативных документов, определяющих структуру и состав ЛО участников сети.

Стандартизация — это разработка единых стандартов, позволяющих произвести унификацию отдельных элементов БО, ИПЯ, терминологии.

Универсальные (общесистемные) языки должны обеспечить единообразие формирования информационных массивов. Примером создания универсальных языковых средств является разработка Государственного рубрикатора научно-технической информации (ГРНТИ).

Метод конверсии, то есть преобразование записей на одном информационно-поисковом языке в записи на другом ИПЯ автоматизированными средствами, реализуется созданием таблиц соответствия. Например, в отраслевом рубрикаторе Центральной научной сельскохозяйственной библиотеки (ЦНСХБ) каждой рубрике Рубрикатора приписан индекс УДК.

Сосуществование языковых средств предполагает параллельное использование нескольких ИПЯ в одной ИПС. Анализ 10 важнейших библиотечных процессов (комплектование, учет библиотечных фондов, библиографическое описание произведений печати, систематизация (или предметизация), организация библиотечного каталога, техническая обработка документов, работа с фондом, обслуживание читателей, работа МБА, справочно-библиографическая и информационная работа) показывает, что ИПЯ в той или иной степени используются в каждом из перечисленных процессов, кроме того, существует прямая зависимость между качеством лингвистических средств и эффективностью используемой библиотечно-библиографической технологии. Следовательно, изменение или расширение функций автоматизированной библиотечной системы связано в первую очередь с реальным выбором комплекса ИПЯ, усилением семантической силы используемых информационно-поисковых языков.

Исследователи отмечают, что, несмотря на существенные достижения в области интерактивных систем (генерация БД, возрастание скорости передачи информации), совершенствование и упрощение поисковой процедуры достигнуто лишь в части автоматизации механических, рутинных процессов интерактивного поиска. Что касается связанных с ним интеллектуальных процессов, то они автоматизацией охвачены слабо или фактически не охвачены. Другими словами, интерактивный поиск дает быстрые результаты по поиску по простейшим элементам базы обслуживания (БО): автору, названию, но тематический поиск, который является интеллектуальным, остается слабым звеном. В исследованиях по анализу эффективности работы интерактивных систем отмечено, что наибольшее влияние на результаты поиска оказывают именно интеллектуальные операции: определение предмета, области поиска, выбор базы данных, выбор стратегии поиска и оценка его результатов. Причем основная сложность заключается в выборе стратегии поиска, что напрямую связано с использованием лингвистических средств. В интерактивном режиме существует задача оптимизации методов поиска, его полноты, релевантности и скорости создания поискового предписания.

ЛО гарантирует формализованное описание содержания документов в ЭК и информационных запросов, что достигается при помощи комплекса ИПЯ. Классификационные и дескрипторные языки служат инструментом более тонкого анализа для проведения тематического поиска. Сочетание нескольких ИПЯ дает возможность проведения поиска по тематическим признакам, что обеспечивает его полноту и точность.

В ЦНСХБ используются для автоматизированного поиска: язык библиографического описания (ЯБО); язык ключевых слов (ЯКЛ); информационно-поисковый тезаурус (ИПТ); отраслевой рубрикатор, разработанный на основе ГРНТИ (ОР).

Результативность поиска в ЭК во многом зависит от выбора стратегии поиска; от лингвистических средств, используемых в данном ЭК; от качества индексирования документов на используемых в электронных каталогах ИПЯ. Семантическая обработка документа подразумевает полноту и точность перевода с естественного языка на ИПЯ, которые зависят от структуры, лексической наполненности и других возможностей информационно-поискового языка, разработанности правил этого перевода, от соответствия единиц естественного языка лексическим единицам ИПЯ. Именно от точности и единообразия описания исходной информации языковыми средствами зависит релевантность (степень соответствия содержания документа, найденного при поиске, содержанию информационного запроса) и полнота поиска.

Если известны источники и реквизиты документа, то поиск ведется по ЯБО, если нужен тематический поиск, то используются ОР, ИПТ, ЯКС.

В ИПС нашей библиотеки используется коммуникативный формат RUSMARC. Структура ЯБО богата поисковыми возможностями, заложенными в этом формате на БО, состоящем из 229 элементов данных. Эти данные позволяют идентифицировать и разыскать документ по каждому из этих элементов. Чем полнее используются возможности коммуникативного формата, тем шире возможности поиска по формальным признакам документа.

Установлено, что поиск только по БО может быть достаточно эффективен, так как заглавия пригодны для автоматизированного поиска. Эффективность поиска возрастает, когда к БО добавляются рубрики или индексы ИПЯ. Точность поиска в этом случае составляет 70 процентов, а полнота — 50 процентов. Точность поиска возрастает еще на 3—5 процентов, если к этому добавляются ключевые слова и дескрипторы. БД с рефератами и/или аннотациями дает максимально эффективный поиск в автоматизированном режиме, поскольку возможен поиск по всем полям, то есть по всему тексту документа. Использование всех текстов документа (БО, аннотаций, рефератов) в качестве ПОД расширяет возможности поиска, так как в них выражены синтаксические связи между ключевыми словами.

Результативность тематических запросов зависит от ИПЯ, на котором они сформулированы. Запрос может быть сделан на естественном языке, то есть выражен известными пользователю терминами — научными или общеупотребительными, и какое-то количество нужных пользователю документов может быть найдено. Однако, как показал опыт, это будут не все документы по заданной теме и, возможно, в выборку не

войдут самые ценные из них, о чем пользователь может и не подозревать. Может показаться, что поисковые возможности естественного языка и ключевых слов одинаковы, но это не так. К примеру:

- в документе № 1 препарат А упоминается в качестве стандарта при оценке свойств препарата Б;
- в документе № 2 описаны свойства, формы, назначения, способы применения и т. п. препарата А.

На запрос <препарат А> и при поиске по текстовым полям (естественней язык) пользователь получит оба документа, так как в их текстовых полях, например, в аннотации, в реферате, поисковая система найдет термин <препарат А>. Однако документ № 1 не релевантен запросу и не нужен пользователю (это «информационный шум»). Документ № 1 релевантен только запросу о препарате Б. На запрос <препарат А> и при поиске по терминам поля «ключевые слова» поисковая система выдаст только релевантный запросу документ № 2, поскольку индексатор заиндексировал документ ключевым словом *препарат А*, так как в нем содержится существенная информация об этом препарате, в отличие от документа № 1.

Но следует иметь в виду, что поиск по терминам текста и ключевым словам не может обеспечить удовлетворительной полноты нахождения нужных источников информации. К примеру, если препарат А в документах № 1 и № 2 имеет разные наименования, что очень распространено в научных текстах. При этом версии написания термина, использованные в документе и, следовательно, индексатором в качестве текста ключевых слов могут отличаться от версии термина, использованной пользователем в запросе. Очевидно, что в таких случаях поисковая система не найдет значительное количество документов. В числе недополученных пользователем могут быть особенно ценные и релевантные его запросу документы.

В качестве ключевых слов (КС) могут выступать отобранные из текста документа слова или словосочетания естественного языка, раскрывающие наиболее важные смысловые аспекты документа. Для пользователя поиск будет наиболее эффективным, если формулировка его запроса совпадет с дескрипторами ИПТ. ИПТ представляет собой алфавитный перечень отраслевой терминологии, где отражены иерархические, синонимические и ассоциативные отношения между терминами (дескрипторами).

Использование дескрипторов ИПТ позволяет систематизатору преодолеть такую особенность естественного языка, как неоднозначность (одно и то же понятие может быть сформулировано поразному), а всем специалистам в данной области — единообразно переводить слова естественного языка на ИПЯ. Это повышает вероятность того, что пользователь сможет найти данный документ. Благодаря тезаурусу, при поиске пользователь может использовать в запросе синонимы, в то же время в тезаурусе есть отсылка от синонима к основному термину, то есть документ все равно будет найден по основному термину.

При индексировании документов КС индексатор в целях обеспечения полноты отражения понятий и релевантности поиска выбирает именно дескрипторы ИПТ, однако бывает, что используемый автором исходного документа термин является очень узким и специфичным либо редко встречающимся в специальной литературе, и поэтому, естественно, что он еще не нашел отражения в ИПТ. В этом случае индексатор может отразить понятие в виде КС, которое считает оптимальным. Понятно, что термины ИПТ все индексаторы напишут одинаково, а формулировки прочих КС теоретически могут оказаться не совпадающими, и тогда пользователь не сможет найти часть документов. Для предотвращения подобных ситуаций в нашей библиотеке постоянно ведется работа по унификации написания КС, составляется картотека КС, фиксирующая согласованные формулировки КС, принимаются методические решения, которые заносятся в специальные рекомендации по индексированию на языке ключевых слов.

Тезаурус и ключевые слова дают эффективный узкотематический поиск. Важное значение имеет использование методик индексирования на ИПЯ, используемых в ИПС. Методики способствуют унификации индексирования документов, препятствуют проявлению субъективизма индексатора в определении места документа, обеспечивают точность, полноту и однозначность отображения информации в БД.

Индексирование — это основное средство раскрытия содержания документа и соответственно всего текущего документного потока, который составляет фонд библиотеки. От качества индексирования зависит не только эффективность тематического поиска в информационных ресурсах, но и эффективность использования ее фондов.

Независимо от типа ИПЯ основными требованиями, которые предъявляются к процессу индексирования документа, являются: а) полнота и точность раскрытия содержания; б) объективность его раскрытия; в) единообразие отображения средствами данного ИПЯ сходных по содержанию документов (другими словами все документы по одному вопросу должны получить одинаковые индексы, рубрики, дескрипторы и т. д. и попасть в одно место в информационно-поисковой системе).

Процесс индексирования включает несколько этапов: анализ содержания документа; выявление и отбор понятий, тем, отражающих основное содержание документа; выбор терминов индексирования (рубрик, кодов, индексов, дескрипторов, ключевых слов) и принятие решений о составе ПОД; перевод содержания документа с естественного языка на ИПЯ; добавление любой необходимой информации к названию документа (расширение названия, создание аннотации); редактирование терминов индексирования на ИПЯ.

Как для классификационных (УДК, ББК), так и для дескрипторных (тезаурус) ИПЯ полнота и детальность индексирования связаны с обеспечением полноты и релевантности тематического поиска.

Полнота и детальность индексирования зависят от семантической наполненности ИПЯ, его способности описать документ в характеристиках, присущих индексируемому документу. Повышение глубины (детальности) индексирования увеличивает точность информационного поиска, его эффективность за счет возможности предоставления информации по самым «узким», специальным вопросам.

Поэтому при создании автоматизированной ИПС, электронного каталога библиотека стоит перед выбором лингвистического обес-

печения, которое будет в них использоваться.

Состав и структура ЛО автоматизированной системы связаны с функциями библиотеки. От выбора ИПЯ и лингвистических средств зависит эффективность работы ИПС. При выборе ЛО необходимо учитывать тематический диапазон фонда, отрасль знаний, представленную в фонде и информационных ресурсах, структуру и объем входного документного потока, тип и особенности ИПС, информационные запросы пользователей. Именно задачи, стоящие перед ИПС, определяют выбор и состав лингвистических средств, совокупность которых должна обеспечить ее эффективную работу.

Оптимизация структуры лингвистического обеспечения автоматизированной ИПС заключается в формировании структуры, которая включает информационно-поисковые языки, обеспечивающие все ее библиотечно-библиографические процессы и функции как на внутрибиблиотечном, так и на межбиблиотечном уровне. Лингвистические средства ИПС должны обеспечивать эффективный информационный поиск. Это могут быть ИПЯ, специально разработанные для автоматизированных ИПС, либо приспособленные для работы в них. Для формирования структуры лингвистического обеспечения ИПС нашей библиотекой разработана методика, которая может быть применена при формировании ЛО научных сельскохозяйственных библиотек и библиотек других ведомств.

Методика формирования структуры лингвистического обеспечения ИПС включает несколько этапов:

1) анализ задач, стоящих перед библиотекой, ее функций и библиотечно-библиографических процессов. Задачи определяют функции, которые реализуются технологиями. Выявление библиотечно-библиографических процессов позволяет определить лингвистические средства, требуемые для их обеспечения;

2) изучение роли и функций ИПЯ в ИПС. Ознакомление с теорией лингвистического обеспечения позволяет понять назначение и роль ИПЯ в формировании и структурировании информационных массивов, в аналитико-синтетической обработке информации, информационном поиске и т. д.

3) анализ эффективности использования собственных ИПЯ позволяет понять, как уже используемые в библиотеке информационно-поисковые языки обеспечивают автоматизированные библиотечно-библиографические процессы, наметить пути совершенствования и адаптации их к автоматизированной ИПС;

4) изучение существующих отраслевых ИПЯ. В случае, если собственные ИПЯ не обеспечивают эффективное функционирование ИПС (эффективный информационный поиск), изучение структуры и поисковых возможностей, методических пособий отраслевых и других ИПЯ позволит определить, подходят ли они данной ИПС;

5) создание структуры лингвистического обеспечения: подбор ИПЯ, определение функций каждого ИПЯ в структуре с учетом внутрибиблиотечных процессов и существования библиотеки в едином информационном пространстве отрасли;

6) адаптация выбранных лингвистических средств к условиям ИПС: проведение работ, обеспечивающих использование ИПЯ в ИПС и выполнение правил работы с ними, усовершенствование ИПЯ с целью повышения эффективности их использования, разработка методических пособий.

Создание ЭК и БД потребовало разработки специальных ИПЯ, приспособленных для автоматизированного поиска, которые также входят в структуру ЛО ИПС ЦНСХБ.

Модель структуры ЛО должна основываться на практической значимости и научной обоснованности ценности каждого ИПЯ в ИПС. Применение ИПЯ, которые не используются в автоматизированной системе, может быть оправдано только их использованием для другого рода тематического поиска.

Модель структуры ПО нашей библиотеки состоит из двух уровней: внутрибиблиотечного значения и межбиблиотечного значения, где субъекты ЛО реализуют свои функции.

Разноуровневая модель структуры ЛО ЦНСХБ раскрывает взаимодействие, функции ИПЯ в ИПС ЦНСХБ. Модель позволяет определить функциональную «нагруженность» каждого ИПЯ в зависимости от нарастания или убывания функций. Модель структуры ЛО позволяет выявить те ИПЯ, роль которых в автоматизированном информационном поиске возрастает, и наметить пути оптимизации именно этих лингвистических средств.

Структура лингвистических средств ЦНСХБ в соответствии с ее оптимизированной моделью выглядит так:

внутрибиблиотечный уровень:

- ЯБО (для идентификации документов и информационного поиска по полям коммуникативного формата);
- УДК (для индексирования входного документного потока);
- ОР (для индексирования входного документного потока и тематического поиска в БД; структурирования информационных массивов; формирования текущих библиографических и реферативных изданий; определения тематического диапазона библиотечных фондов ЦНСХБ);
- ИПТ (используется для индексирования входного документного потока и тематического поиска в БД; создания терминологической базы по сельскому хозяйству и продовольствию);
- ЯКС (для индексирования входного документного потока и тематического поиска в БД; отбора лексики в информационно-поисковый тезаурус по сельскому хозяйству и продовольствию);

межбиблиотечный уровень:

- УДК (в корпоративной каталогизации и АСОД, а также в качестве международного информационного языка);
- ЯБО (в корпоративной каталогизации и АСОД и для идентификационного поиска информации в БД страны);
- ОР (как язык-посредник межотраслевого информационного общения, для обмена информацией и ее поиска в ИПС РФ и других стран СНГ, а также в качестве общеотраслевого ИПЯ АПК);
- ИПТ (как терминобаза АПК, а также в качестве общеотраслевого ИПЯ АПК).

На примере оптимизации структуры лингвистических средств Центральной научной сельскохозяйственной библиотеки видно, что составе ее лингвистических средств целесообразно оставлять только те информационно-поисковые языки, которые будут использоваться в автоматизированном поиске.