

3. Самохіна, Н. Ф. Формування наукової електронної бібліотеки Національної бібліотеки України імені В. І. Вернадського / Н. Ф. Самохіна // Наук. пр. / Нац. б-ка України ім. В. І. Вернадського. – Київ, 2012. – Вип. 34. – С. 13–17.

ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ В НАУЧНЫХ БИБЛИОТЕКАХ

Л. Н. Пирумова

Центральная научная сельскохозяйственная библиотека, г. Москва,
Россия

Лингвистическое обеспечение (ЛО) является сегодня неотъемлемой частью автоматизированной системы, а впервые о его необходимости заговорили в 1960-е годы, когда стало понятно, что без структурирования любой набор информации остается всего лишь сундуком с множеством беспорядочных данных, а компьютеры — просто железными ящиками, набором железа. Только информационно-поисковый язык (ИПЯ) превращает горы собранной информации в структурированную, организованную систему, способную не только хранить в определенном порядке, но и выдавать нужную информацию. Тогда стали много и часто говорить о важности и сложности разработки ЛО, стала развиваться теория информационного поиска. А с появлением Интернет опять горячие головы стали говорить, что ЛО не нужно, все можно найти просто в Интернет. Однако и эта волна пренебрежения к ЛО затихает, опять говорят о его важности уже для поиска в Интернет.

Понятие ЛО менялось с развитием и изменением состава и функций информационно-поисковых систем (ИПС). На начальных этапах развития автоматизированных систем под этим понятием понимали комплекс ИПЯ, используемых для создания поискового образа документа и поисковых предписаний, позже ЛО стало трактоваться как совокупность ИПЯ и нормативных документов, определяющих правила их ведения и использования. С конца 1980-х годов ЛО трактуется как комплекс средств для предоставления информации в виде данных и интерпретации данных в процессе

функционирования автоматизированных ИПС. Ключевым элементом комплекса является ИПЯ — искусственный язык, предназначенный для формализованного описания смыслового содержания документов и последующего их поиска. Различают:

1) ЛО процессов подготовки баз данных (БД), которое в свою очередь состоит из ЛО аналитико-синтетической обработки документов, ЛО процессов ввода и корректуры текстов документов, ЛО формирования БД;

2) ЛО подготовки информационных изданий, включая их структурирование и полиграфическое оформление;

3) ЛО процессов распределения потоков информации между потребителями: ЛО распределения информации; ЛО широкотематического и узкотематического поиска, общесистемные языковые средства общения между ИПС.

Противники использования ИПЯ в информационном поиске говорят, что делать запрос на естественном языке легче, чем на искусственном. Но поскольку ИПЯ специально разработаны для автоматизированного поиска, то они лишены недостатков естественного языка с точки зрения формирования информационных запросов и поисковых предписаний: избыточности (наличие предлогов, союзов, междометий, которые не несут смысловой нагрузки), неоднозначности выражения. Использование в естественном языке синонимов значительно украшает естественный язык, но приводит к неоднозначности обозначения понятия в информационной системе — оно будет расположено в разных ее местах. Наличие полисемии, омонимии также усложняет поиск на естественном языке, создает информационный шум и не обеспечивает точность поиска. ИПЯ лишены этих недостатков поскольку соответствуют требованиям, важным для автоматизированного поиска: однозначность (каждая запись на ИПЯ должна иметь только один смысл); наличие семантических, логических отношений и психологических ассоциаций; возможность корректировки и дополнения новыми терминами, понятиями при сохранении структуры; удобство пользования; способность идентифицировать предмет. Семантическая сила, богатство ИПЯ выражается его терминологической базой, разнообразием парадигматических отношений, развитостью структуры. Наиболее сложная составляющая ИПЯ — разработка отношений между терминами,

которые являются основой языка. Парадигматические отношения не зависят от контекста документа, характеризуются отношениями тождества (отношения предпочтительности в использовании синонимии), иерархии (отношения подчинения типа «род-вид», частное-целое), ассоциации (отношения «причины-следствия», смежности, контраста, кратности).

ИПЯ является средством свертывания информации в процессе смысловой обработки документов, в процессе их аналитико-синтетической обработки, средством их формализованного описания. Поэтому ИПЯ тесно связаны с семантической обработкой документов, для которой они создаются и в процессе которой они пополняются новой терминологией. Как средство семантической обработки ИПЯ должен быть удобен для работы с ним индексатора (систематизатора) и обладать компактностью записей, способствующих его запоминанию. ИПЯ должен обеспечить полноту и точность поиска, для этого он должен обеспечить возможность индексатору точно идентифицировать предмет и описать его особенности с необходимой полнотой и детализацией.

Очевидно, что создание ИПЯ — процесс трудоемкий, среди научных сельскохозяйственных библиотек России только федеральное государственное бюджетное научное учреждение «Центральная научная сельскохозяйственная библиотека» (ЦНСХБ) занимается их разработкой. ЦНСХБ создает Отраслевой рубрикатор на основе Государственного рубрикатора научно-технической информации (ГРНТИ) и Информационно-поисковый тезаурус по сельскому хозяйству и продовольствию, а также разрабатывает Отраслевые рабочие таблицы УДК для АПК на основе эталонных таблиц УДК на русском языке, создаваемых ВИНТИ РАН. К каждому из ИПЯ разработаны свои методические пособия. Разработка ИПЯ в ЦНСХБ проходит несколько этапов: отбор лексических единиц; процесс нормализации лексики; систематизация и группировка лексики; построение классификационных схем; словарных статей, оформление лексики ИПЯ. Этап отбора лексических единиц особенно важен в процессе создания ИПЯ, поскольку от него зависят возможности данного ИПЯ, его терминологическая наполненность, соответствие уровню развития науки, отражаемой в нем, а значит и поисковые возможности данного ИПЯ.

ИПЯ в ЦНСХБ неразрывно связан с процессом аналитико-синтетической обработки информации, поскольку на этом этапе происходит отбор новой лексики в ИПЯ. Прежде чем отобранная терминология, лексика предстанет в виде элементов ИПЯ, она проходит семантическую, то есть смысловую обработку, в процессе которой создаются словарные статьи терминологического понятия, его парадигматические (смысловые) отношения, выявляются синонимы и т. д. ИПЯ, разрабатываемые в ЦНСХБ, имеют электронную версию, поэтому многие технологические процессы ведения ИПЯ делаются в автоматизированном режиме, в том числе корректура, правка терминологии. Ежегодно тезаурус пополняется примерно на 1 500 терминов. В 2014 году его объем превысил 43 тысячи терминов. Кроме того, в качестве приложений к тезаурусу создаются авторитетные файлы по различным областям науки (Фунгициды, Инсектициды и т. д.).

Следуя рекомендациям Н.И. Гендиной [1, с. 168] при исследовании применения ИПЯ в современной научной сельскохозяйственной библиотеке и их анализе использовался технологический подход. Были проанализированы основные технологические процессы в автоматизированной информационно-библиотечной системе ЦНСХБ и выявлено, какие лингвистические средства используются при их осуществлении. В процессе анализа установлено, что лингвистические средства в той или иной степени используются в каждом из основных технологических процессов. В структуру лингвистических средств, обеспечивающих функционирование автоматизированной информационно-библиографической технологии входят 5 ИПЯ: язык библиографического описания (формат RUSMARC) (ЯБО); Отраслевой рубрикатор (ОР), созданный на основе ГРНТИ; Информационно-поисковый тезаурус по сельскому хозяйству и продовольствию (ИПТ); а также Универсальная десятичная классификация (УДК) и язык ключевых слов (ЯКС).

Использование нескольких ИПЯ в одной ИПС объясняется тем, что каждый из языков предназначен для выполнения определенных функций в ней, а также тем, что нет и пока не может быть создан единый ИПЯ, выполняющий одновременно все функции лингвистических средств и обеспечивающий все задачи, стоящие перед ИПС. Одновременное использование нескольких ИПЯ 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.

В свое время для формирования структуры ЛО ИПС в ЦНСХБ была разработана методика, по которой создана модель структуры научной сельскохозяйственной библиотеки. Модель структуры ЛО ЦНСХБ состоит из двух уровней: уровень внутрибиблиотечного значения и уровень межбиблиотечного значения.

Разноуровневая модель структуры ЛО ЦНСХБ раскрывает взаимодействие, функции ИПЯ в ИПС. Модель позволила определить функциональную «нагруженность» каждого ИПЯ в зависимости от количества функций, выявить те ИПЯ, роль которых в автоматизированном информационном поиске возрастает, наметить пути оптимизации именно этих лингвистических средств и отказаться от тех ИПЯ, функции которых дублировались другими ИПЯ. Так, в результате этого анализа выяснилось, что терминологическая база ИПТ практически равна терминологической базе карточного Комплексно-системного каталога (КСК), а БД «АГРОС» обеспечивает такой же глубокий и разнообразный тематический поиск как КСК, поскольку так же как и он включает наряду с информацией о книгах, информацию о статьях из отечественных и зарубежных периодических и продолжающихся изданий. В результате было принято решение о закрытии карточного каталога для текущего тематического поиска, теперь текущий поиск осуществляется только в БД «АГРОС», а КСК используется для ретропоиска.

Оптимизированная модель структуры ЛО автоматизированной ИПС ЦНСХБ, выглядит так:

Внутрибиблиотечный уровень:

- ЯБО (для идентификации документов и информационного поиска по полям коммуникативного формата);
- УДК (для индексирования входного документального потока, организации фондов открытого доступа);
- ОР (для индексирования входного документального потока и тематического поиска в БД; структурирования информационных массивов; формирования текущих библиографических и реферативных изданий; определения тематического диапазона библиотечных фондов ЦНСХБ);

- ИПТ (для индексирования входного документального потока и тематического поиска в БД; создания терминологической базы по сельскому хозяйству и продовольствию);

- ЯКС (для индексирования входного документального потока и узкотематического поиска в БД; отбора лексики в информационно-поисковый тезаурус по сельскому хозяйству и продовольствию).

Межбиблиотечный уровень:

- УДК (в корпоративной каталогизации и индексировании входного документального потока, а также в качестве международного информационного языка);

- ЯБО (в корпоративной каталогизации и индексировании входного документального потока, для идентификационного поиска информации в БД страны);

- ОР (как язык-посредник межотраслевого информационного общения, для обмена информацией и ее поиска в ИПС России и СНГ, а также в качестве общеотраслевого ИПЯ АПК);

- ИПТ (как терминобаза АПК, а также в качестве общеотраслевого ИПЯ АПК).

В этой структуре каждый ИПЯ выполняет свою задачу и функции, обеспечивая разнообразную деятельность библиотеки. Поскольку в ЦНСХБ используются несколько ИПЯ, то встает вопрос об их совместимости. Для того, чтобы добиться совместимости лингвистических средств в ИПС, используются методики индексирования для каждого из ИПЯ, а также стандартизация языковых средств и поддерживающих компонентов ЛО.

Используется также такой метод достижения совместимости как создание таблиц соответствия. Например, в ОР каждой рубрики ОР приписан индекс УДК. Кроме того, создаются микротезаурусы по отдельным отраслям АПК, на базе которых создаются таблицы соответствия УДК и ОР, то есть к дескриптору микротезауруса приписывается индекс УДК и рубрика ОР.

Сложнее добиться совместимости ЛО библиотек АПК России. В эпоху карточных каталогов библиотеку обязывали соответствующими постановлениями правительства СССР строить систематический каталог либо по ББК (общественные науки), либо по УДК (технические и естественные науки), обеспечивая таким образом совместимость лингвистических средств разных

библиотек, а пользователю — легкий переход из поисковой системы (каталога) одной библиотеки в другую. Поскольку выбор ИПЯ в настоящее время не регламентирован, и каждая библиотека свободна в выборе ИПЯ, то обмен информацией между ними может быть затруднен из-за несовместимости этих ИПЯ. В настоящее время сеть научных сельскохозяйственных библиотек России включает библиотеки: сельскохозяйственных учебных заведений (в т. ч. 58 библиотек вузов и 11 учреждений дополнительного образования), научно-исследовательских учреждений (168), опытных станций (16), а также 2 библиотеки со статусом центральной — ЦНСХБ и Сибирская научная сельскохозяйственная библиотека. Для достижения совместимости ЛО ИПС АПК ЦНСХБ предлагает использовать единые общепрофессиональные ИПЯ. Такими общепрофессиональными ИПЯ являются: Отраслевой рубрикатор, Отраслевой Информационно-поисковый тезаурус по сельскому хозяйству и продовольствию, дескрипторный словарь по сельскому хозяйству и продовольствию, а также общероссийский библиотечный формат записи (RUSMARC). Использование в структуре ЛО научной сельскохозяйственной библиотеки общепрофессиональных ИПЯ способствует решению проблемы совместимости ИПЯ в ИПС отрасли, которая является достаточно актуальной, и созданию единого информационного пространства отрасли. Необходимо также обеспечить методологическую совместимость, стандартизацию и унификацию языковых средств.

Казалось бы, значительно проще методически обеспечить совместимость, то есть принять (разработать) единые принципы ведения ЛО отдельных ИПС, входящих в информационную систему сельскохозяйственных библиотек отрасли (разработать нормативные документы, определяющих структуру и состав ЛО участников сети), однако пока не удалось принять такой единый документ. В этих условиях единственным способом, обеспечивающим совместимость ИПС АПК, является использование общепрофессиональных ИПЯ, которые должны обеспечить единообразие формирования информационных массивов. В этом направлении есть подвижки, но не все библиотеки готовы к освоению новых для них лингвистических средств. С 2011 года ЦНСХБ создает Сводный каталог фондов научных библиотек АПК, который строится на автоматизированных программных и лингвистических средствах ЦНСХБ. В основе проекта лежит

современная информационная технология Облачных вычислений (Cloud Computing) для максимального упрощения и удешевления процесса создания собственных каталогов библиотек-участниц и автоматизации работы этих библиотек. Проект обеспечивает создание: электронного сводного каталога фондов библиотек АПК; электронных каталогов фондов библиотек-участниц; системы книговыдачи в каждой библиотеке; объединенной службы МБА; сводной электронной библиотеки научных учреждений АПК. Разработка опирается на международные стандарты Интернет, Российский коммуникативный формат представления библиографических записей (RUSMARC), Российские правила каталогизации. На апрель 2014 года в Сводном каталоге зарегистрировано 160 библиотек, из них 38 активно ведут свои собственные электронные каталоги. Для поиска в Сводном каталоге доступно 309 266 документов. Развитие этого проекта будет способствовать освоению научными библиотеками общепрофессиональных ИД.

Проблема совместимости лингвистических средств существует не только в российских, но также в международных и зарубежных БД по вопросам АПК, и решается она разными методами. Наиболее авторитетными БД, охватывающими весь спектр вопросов сельского хозяйства, пищевой промышленности и проблем, смежных с ними и содержащими основной объем информации (около 65 %) по сельскому хозяйству, являются 3 БД: «AGRICOLA» (National agricultural library (NAL) — Национальная сельскохозяйственная библиотека США), БД «AGRIS» (Food Agricultural Organization (FAO) — Международная информационная служба по сельскому хозяйству ФАО) и «CAB ABSTRACTS» (Commonwealth Agricultural Bureaux International (CABI) — Международное сельскохозяйственное бюро стран Британского содружества). Все эти зарубежные БД используют в качестве лингвистических средств рубрикаторы — коды тематических рубрик и тезаурусы. Рубрикаторы используются, главным образом, для формирования реферативных журналов, формирования крупных информационных массивов и при неглубоком тематическом поиске.

Наиболее интересны усилия по совместимости тезаурусов этих БД. БД «AGRIS» создается усилиями более чем 100 стран, которые поставляют в нее информацию о национальных документах своих

стран, поэтому тезаурус AGROVOC многоязычный, существуют версии на 18 языках, что для стран-участников решает проблему совместимости ЛО их БД с БД AGRIS. В 2011 году ЦНСХБ разработала русскоязычную версию этого тезауруса объемом более 32,9 тысяч терминов, эквивалентных терминам тезауруса AGROVOC.

CABI и NAL одно время решали проблему совместимости, используя общий тезаурус: с 1985 по 2005 гг. в БД «AGRICOLA» использовался тезаурус CABI, что объяснялось не только трудоемкостью создания собственно тезауруса, но и общей тематикой этих БД и общим естественным языком — английским. Однако даже при этом существовала проблема национальной терминологии, которая решалась в этом тезаурусе включением в него дескрипторов, которые используются только в БД «AGRICOLA». Однако затем NAL разработала собственный тезаурус. В конце 1990-х гг. возникла идея создания общего для БД «AGRIS», БД «AGRICOLA» и «CAB ABSTRACTS» тезауруса, которая неоднократно обсуждалась специалистами-создателями тезаурусов в CABI и AGRIS. Были разработаны идеология, проект и структура такого тезауруса. Известно, что он должен был состоять из 2-х частей: алфавитной и систематической. Была создана группа разработчиков нового тезауруса, но из-за финансовых проблем этот проект не был реализован. Однако в начале 2014 года к этой идее вновь вернулись, и она обсуждается в сообществе ФАО. Не менее важной является совместимость лингвистических средств отечественных и зарубежных БД по сельскому хозяйству, но пока в этом направлении сделано очень немного. Из российских БД по объему и тематическому охвату с ними может сравниться только БД ЦНСХБ «АГРОС». В целях разработки мер по обеспечению совместимости тезаурусов этих БД и тезауруса ЦНСХБ в свое время было проведено сравнительное изучение этих тезаурусов, которое позволило сделать некоторые практические выводы для усовершенствования тезауруса ЦНСХБ и повышения его совместимости с зарубежными тезаурусами.

Прежде всего ИПЯ создаются для информационного поиска. В БД «АГРОС», которая является основным информационным продуктом ЦНСХБ, используются для автоматизированного поиска все описанные ранее ИПЯ, кроме УДК. Результативность поиска в БД во многом зависит от выбора стратегии поиска; от

лингвистических средств, используемых в ней; от качества индексирования документов на них. 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. Выбор ИПЯ ЦНСХБ объясняется тем, что именно эти лингвистические средства обеспечивают качественный информационный поиск, что было подтверждено различными исследованиями, в том числе и в самой ЦНСХБ.

Структура ЯБО богата поисковыми возможностями, заложенными в формате RUSMARC, 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. Этот ИПЯ обеспечивает идентификационный поиск, когда известны источники и реквизиты документа. При тематическом поиске используются ОР, ИПТ, ЯКС. Эффективность поиска возрастает, когда к библиографическому описанию добавляются рубрики ОР. Точность поиска в этом случае составляет 70 %, а полнота — 50 %. Точность поиска возрастает еще на 3–5 %, если к этому добавляются дескрипторы ИПТ. Исследования, проведенные в ЦНСХБ, подтверждают эффективность и преимущества тематического поиска по ИПТ.

Запрос, сформулированный на естественном языке, может быть результативным, то есть пользователь получит некоторое количество документов. При этом, если он не знает, что он конкретно может получить, у него будет иллюзия, что его запрос удовлетворен полностью. Однако будут найдены не все документы по заданной теме и, возможно, в выборку не войдут самые ценные из них, о чем пользователь может и не подозревать, поскольку версии написания термина, используемые пользователем в запросе, могут отличаться от версии написания термина, использованной в документе и, соответственно, индексатором в качестве ключевых слов. Очевидно, что в таких случаях поисковая система не найдет значительное количество документов, и пользователь не получит исчерпывающе полной информации по своему запросу. В числе недополученных документов могут быть именно особенно ценные и релевантные его запросу документы. Возможен и другой вариант:

поиск по полному тексту дает избыточную информацию, только упомянутую в документе, не являющуюся его основной темой.

Использование дескрипторов ИПТ позволяет систематизатору преодолеть такую особенность естественного языка, как неоднозначность (одно и то же понятие может быть сформулировано по-разному), 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, но, поскольку в тезаурусе есть отсылка от синонима к основному термину, документ все равно будет найден по основному термину.

В тех библиотеках, где тезаурус не является средством тематического поиска, а в качестве такового используется ЯКС, предлагается использовать тезаурус или дескрипторный словарь, созданный на его основе в качестве нормализованного словаря отраслевой лексики, обеспечивающего унифицированное представление терминологии в ИПС научной библиотеки.

Большое значение имеет использование методик индексирования на ИПЯ, используемых в научной библиотеке. Методики способствуют унификации индексирования документов, гарантируют соотнесение документов одной тематики в одно, определенное методикой индексирования место в структуре ИПЯ; препятствуют проявлению субъективизма индексатора в определении места документа, обеспечивают точность, полноту и однозначность отображения информации в БД.

Таким образом, с созданием и развитием общеотраслевых лингвистических средств, соблюдением методов совместимости появляется реальная возможность создания единого информационного пространства отрасли.

Список использованных источников:

1. Гендина, Н. И. Лингвистическое обеспечение автоматизированных библиотечных систем / Н. И. Гендина. — Алматы : Гылым, 1991. — 222 с.
2. Пирумова, Л. Н. Лингвистическое обеспечение информационно-поисковых систем библиотек АПК: методика формирования и пути совершенствования : автореф. дис. ... канд. пед. наук : 05.25.03 /

Л. Н. Пирумова ; Моск. гос. ун-т культуры и искусств. – М., 2003. – 19 с.

ИНФОРМАЦИОННАЯ СИСТЕМА ИНТЕРНЕТ- МОНИТОРИНГА ПУБЛИКАЦИЙ: ФУНКЦИИ И СТРУКТУРА

С. Ф. Липницкий, Л. В. Степура, А. Г. Буравкин
Объединенный институт проблем информатики НАН Беларуси,
г. Минск, Беларусь

Введение

В настоящее время в связи с интенсивным увеличением объемов текстовой информации, представленной в электронном виде, все более необходимой становится разработка систем, обеспечивающих решение широкого круга задач мониторинга и аналитической обработки таких данных. При этом постоянно повышаются требования к эффективности процессов поиска текстов и формирования отчетных документов по результатам мониторинга. Все это существенно усложняет получение высоких характеристик работы информационных систем. Другими словами, результаты мониторинга должны отвечать приемлемым требованиям полноты и точности.

В существующих системах автоматической обработки текстовой информации используются подходы по увеличению эффективности поисковых механизмов, которые основаны главным образом на различных лингвистических и статистических методах, что не позволяет достичь требуемого качества работы данных систем. Значительными проблемами для современных информационных служб являются адаптация к информационным запросам конкретных пользователей, а также необходимость проведения углубленного мониторинга по интересующим их тематикам.

Целью данной работы является определение функций и разработка структуры информационной системы Интернет-мониторинга публикаций, в которой настройка на различные предметные области и создание соответствующих баз знаний будет