Информационно-поисковые языки

Пименов Е.Н., Левашова Л.Г., Захаров В.П.

Принципы разработки тезауруса по проблемам сохранности документов

1. Процесс разработки документальных систем обычно включает в себя ряд этапов, и на каждой стадии разработки на передний план выдвигается решение какой-то одной полагаемой наиболее важной задачи. В начальный период работы основная задача - приобретение компьютера, технической базы для современных математических и лингвистических средств организации, хранения и многоаспектного поиска информации. Второй этап – это выбор программного обеспечения. Выбор пакета программ для БД, даже если он производится с участием программистов высокой квалификации, часто имеет случайный характер и редко является хорошо обоснованным, так как во многом зависит от финансовой стороны и возможностей разработки определенной документальной системы. Третье – решение вопроса об информационно-поисковом языке (ИПЯ) базы данных. Выбор может осуществляться между, например, систематизационными, предметизационными и дескрипторными языками, и это решение определяется, в частности, общим характером обрабатываемой информации и назначением документальной БД. Четвертый этап наступает тогда, когда разработчики базы данных получают отчетливое представление о том, что главное – это методика индексирования информации и хорошее качество выполнения индексационных работ, которые иногда хорошо компенсируют общеизвестные недостатки пред- или посткоординативных ИПЯ. Имеется также пятый, шестой и, наверное, другие сменяющие друг друга шаги в процессе создания и совершенствования автоматизированных документальных систем, которые, с наших позиций, можно наметить и выстроить, например, таким образом. Главное – это:

- учет при работе с системой всегда специфических информационных потребностей пользователей;
- создание поисковых интерфейсов, учитывающих данное обстоятельство;
- начало работы в сети Интернет и, в связи с этим, возможное уменьшение интереса к созданию и совершенствованию собственных документальных систем;
- приобретение негативного опыта поисков в Интернете и осознание того, что “свои” базы данных при должной их организации могут иметь преимущества по сравнению с Интернетом и т.д.

Рассматриваемая в этой работе система по обеспечению сохранности документов (далее – БД ОСД) в настоящее время находится на 3-4 этапах из перечисленных выше стадий создания документальных систем. Дескрипторный

англо-русский словарь (информационно-поисковый тезаурус - ИПТ) базы данных в настоящее время включает в себя более 5 тыс. лексических единиц (ЛЕ), слов и терминологических словосочетаний, из которых 1437 составляют дескрипторы - классы условной эквивалентности ИПТ и называющие эти классы слова и около 4 тыс. – синонимы этих дескрипторов. Основа тезауруса была сформирована на материале 9,7 тыс. библиографических записей по тематике ОСД, часть из которых (около 2 тыс.) была обработана обычным порядком, то есть вручную, другие прошли обработку с использованием разработанной в БАН программы автоматического индексирования документов. Цель данной статьи – показать, что многие элементы и стадии разработки документальных систем являются тесно взаимосвязанными, и особенно это относится к правилам разработки тезаурусов и имеющимся в каждой БД рекомендациям или методикам индексирования информации.

2. Одна из особенностей словаря по проблемам обеспечения сохранности документов заключается в том, что процесс его составления отчасти основывается на обобщенной структурной модели (моделях) анализа и описания содержания информации, применяемой также (и прежде всего) при индексировании библиографических записей. В работе по индексированию документов, а с другой стороны, в процессе дескрипторизации ЛЕ применяется пятичленная обобщенная схема или модель индексирования S-Attr-P-Instr-Loc, структурные элементы которой интерпретируются следующим образом:

- **S –** предмет индексирования информации. В БД ОСД в этой функции выступают слова, называющие артефакты, вещества, материалы или любые иные объекты, имеющие материальный характер. Наименования операций и идеальных предметов (как теории, концепции, методы и т.п. [1-2]) к этому классу лексических единиц, по определению, не относятся;
- **Attr** - характеристики предметных частей или других элементов рассматриваемой нами структуры. Такие характеристики могут входить либо в словосочетания, как КОЖАНЫЕ ПЕРЕПЛЕТЫ, ИСКУССТВЕННОЕ СТАРЕНИЕ, УНИВЕРСАЛЬНЫЕ БИБЛИОТЕКИ, либо употребляться самостоятельно как прилагательные и причастия-унитермы типа “переносной”, “портативный”, “краснодубный” и некоторые другие;
- **P** – аспекты анализа информации, какими являются прежде всего слова с обобщенным значением ‘операция’ или ‘процесс’ и некоторые другие аспекты. Главный аспект рассмотрения информации в БД ОСД – это ‘обеспечение сохранности’ документов. Значение данного термина при индексировании, как и в естественном языке, требует своего дополнения и смыслового развертывания в нескольких отношениях, а именно указания на то, что именно сохраняется (S), каким образом, где и когда ведутся работы или исследования по обеспечению сохранности документов. В указанном

отношении позиция P является главной в рассматриваемой аналитической модели и из нее семантически "вытекают" другие имеющиеся в ней элементы;

· **Instr** - названия методов, способов и технологий, сообщающие информацию о том, каким образом осуществляется P или S-P. Наименования методов, как показано в [3], обозначают какие-то новые для определенной предметной области операции и до известной степени раскрывают техническую сторону их выполнения. В силу такого характера содержания семантической функции Instr ее конкретные экспликации иногда мало чем отличаются в семантическом плане от содержания P с категориальным значением 'операции', и в определенных контекстах имеет место нейтрализация (снятие) семантической оппозиции 'P/Instr';

· **Loc** - локализаторы места и времени, отвечающие на грамматические вопросы "когда?" (19 век, A.D., Tudor time, Early time, 1980-1995 гг. и т.п.) или "где?", например, в каких странах и регионах, средах, условиях работы, организациях и предприятиях ведутся работы или исследования по обеспечению сохранности документов.

При анализе содержания документов рассматриваемая нами модель используется таким образом, что словам, выделяемым в качестве ключевых в индексируемых записях, в неявном виде приписывается определенная функция или роль из числа перечисленных выше пяти семантических или текстовых функций, как, например, в документах: "Реставрация (P) ветхих (Attr) рукописей (S) в Библиотеке Конгресса (Loc) США (Loc)", "Неразрушающие методы (Instr) контроля (P) при [оценке(P) свойств(P1)] бумаги (S) документов(S)", "Adesivi (S) per il restauro (P) librario (Loc) e d'archivio (Loc)" "Biological methods (Instr) in book (S) restoration (P)". Приведенные выше примеры показывают, что рассматриваемая модель индексирования в документах обычно представлена не целиком, но как правило – фрагментарно, в ее сокращенном и редуцированном виде, то есть какие-то элементы структуры S-Attr-P-Instr-Loc в обрабатываемых текстах могут отсутствовать. В таком случае говорят об особом - нулевом - выражении предметов, аспектов или других элементов описываемой аналитической схемы [4].

При разработке тезауруса та же модель применялась и применяется в настоящее время при формировании дескрипторов – классов условной эквивалентности или словарных статей ИПТ. Поскольку предметами информации в БД ОСД могут являться, как выше указывалось, только слова с предметно-вещественным категориальным значением, содержание понятия 'предмет индексирования' в БД ОСД является более четко и определенно очерченным, чем во многих других базах данных за счет исключения из рассмотрения как самостоятельных классов ЛЕ так называемых идеальных предметов. Последняя общая прагматическая установка (на минимизацию объема тезауруса и упрощение его понятийной структуры) имела своим результатом в конечном итоге выделение только 5 основных (центральных для

данной системы) предметов и, соответственно, столько же семантических схем анализа индексируемой информации. Эти предметы и обобщенные схемы, определяющие логику индексирования, представлены в таблице 1, в которой намечена также фасетная фабула [1] ИПТ по обеспечению сохранности документов.

Предметно-аспектные классы, охарактеризованные в таблице 1, представляют собою нечеткие множества семантически значимых единиц и конструктов – документов, запросов, семантических схем индексирования и словарных статей ИПТ. В силу такой их природы предметно-аспектные классы определяются в какой-то мере искусственно и могут пересекать по их содержанию, когда, например, в документе рассматриваются и вопросы обеспечения сохранности произведений печати, и занятый в выполнении этих работ персонал, библиотечные здания или другие предметы. В таких случаях в индексируемой информации выделяется несколько элементов со значением 'предмет' и обработка ее производится с применением двух или более из указанных схем индексирования. Трудности при работе с предметно-аспектными классами заключаются, помимо всего, также в диффузном, расплывчатом (в некоторых случаях) характере их содержания. Так, трудно отчетливым образом установить специфику класса ФОНДЫ ХРАНЕНИЯ по отношению к содержанию рубрики ДОКУМЕНТЫ И МАТЕРИАЛЫ. Наименования "документы" и "фонды хранения" в языке обрабатываемых документов и в тезаурусе по ОСД связаны родовидовой семантической связью, как DOCUMENTS в.STOCK, поскольку во многих контекстах эти термины предстают как практически равноценные по выражаемым ими значениям. Когда в документе рассматривается, например, "обеспечение сохранности библиотечных фондов", то без просмотра такой информации de visu смысл последнего выражения такой же, как и значение словосочетания "обеспечение сохранности документов". Различия в содержании данных предметов иногда раскрывает аспектная часть информации, в частности то, что обеспечение сохранности фондов, по нашей оценке, значительно чаще предполагает проведение массовой обработки – нейтрализации, дезинфекции фондов и др., чем обработку отдельных произведений печати или других материалов. Такие оттенки в значениях (или в употреблении) рассматриваемых нами предметов в текстах заглавий и аннотаций библиографических записей, обрабатываемых в БД ОСД, выявляются далеко не всегда, и не могут служить семантическим основанием для различения обсуждаемых классов предметов. Наиболее отчетливым образом класс информации ФОНДЫ ХРАНЕНИЯ выделяется в тех ситуациях, когда это понятие определяют аспекты, специфичные только для этого класса и почти не используемые как аспекты другой информации, например, ФОРМИРОВАНИЕ, КОМПЛЕКТОВАНИЕ, УЧЕТ ФОНДОВ ХРАНЕНИЯ и другие аспекты, не характерные для предметного класса

ДОКУМЕНТЫ И МАТЕРИАЛЫ. Последние, наиболее специфичные для определенных классов аспекты представляют собою не только их дистинктивные или дифференцирующие признаки, но такие аспекты дают словесные "ключи доступа" к определенным предметно-аспектным фрагментам документального массива БД, как аспект информации КОНСТРУКЦИЯ в поисковых выражениях запросов – ведет к выдаче информации лишь о технических средствах, аспекты СТРОИТЕЛЬСТВО, АРХИТЕКТУРА, ИНЖЕНЕРНЫЕ СИСТЕМЫ – к информации о помещениях и зданиях документохранилищ и т.п. В указанной семантической роли аспекты (в поисковых выражениях запросов) представляют собой, таким образом, языковые аналоги указателей роли, ранее широко применявшиеся во многих документальных системах. Ориентируя проведение поиска на определенный класс документов, аспекты являются лингвистическим средством обеспечения необходимого соотношения коэффициентов точности и полноты выдаваемой пользователям информации.

Содержание элементов структурной модели S-Attr-P-Instr-Loc (или ее вариантов, представленных в таблице 1 можно рассматривать с разных сторон, интерпретировать, например, таким образом:

- как категории, близкие к наиболее общим фасетам классификации Ш.Ранганатана [5]. Здесь мы имеем в виду такие абстрактные категории Ш.Ранганатана, как МАТЕРИЯ, ЭНЕРГИЯ, ВРЕМЯ и МЕСТО и близость к ним содержания семантических функций S, P и Loc. Различия заключаются в том, что категории Ранганатана не имеют контекстовой или синтагматической природы, а элементы структурных моделей – это функции в определенных контекстах, задаваемых схемами индексирования;
- как разновидность "мешочной грамматики", не имеющей явного выражения в поисковых образах документов. Несмотря на такую особенность выражения синтагматических связей ЛЕ эта грамматика, тем не менее, как показано выше, регламентирует процесс индексирования и так же, как указатели роли, ее можно использовать для повышения точности поисков;
- как концептуальные синтагматические отношения в их понимании, представленном в [6]. Последние отношения принадлежат одновременно и к области синтагматики, то есть линейных связей ЛЕ, и к сфере парадигматических связей понятий, являются более абстрактными, чем указатели роли и, может быть, ближе всего к семантическим падежам Ч.Филлмора [7]. Синтагматичность рассматриваемых 5 категорий проявляется, в частности, в том, что относящиеся к одному "внеконтекстному" классу фасеты получают или могут иметь различные интерпретации в разных схемах анализа информации. Так, ЛЕ со значением 'организации' по отношению к схеме 1, то есть в контекстах как "Анализ сохранности (P) рукописного фонда А.А.Ахматовой (S) в РНБ (Loc)" выступает в позиции

локализатора места, а в контексте и схеме анализа 6 – “IFLA PAC program”, “Деятельность (P) подкомитета по консервации библиотечных фондов (S1) Библиотечной ассоциации (S) Великобритании (Loc)” - такие слова получают значение предметов анализа информации;

· как модель, изоморфная по своему содержанию, структуре главных-второстепенных членов предложения и “накладываемая” на индексируемый текст с целью обеспечения простоты и единообразия его индексирования [4]. Явно выраженная прагматичность последнего понимания значения и назначения рассматриваемой нами структуры имеет своим преимуществом то, что с этих позиций не очень существенно относятся ли эти функции (и структуры) в области онтологии, представлены в обрабатываемых документах и в их языке или же эти понятия – метаязыковые и относятся только к ИПЯ, которые, как известно, имеют характер отчасти искусственно сформированных лингвистических конструктов [8-9]. С наших позиций этот вопрос не имеет большого значения, если использование рассматриваемых схем информации дает ощутимые положительные результаты в ходе ее обработки.

Не входя здесь в дальнейшее рассмотрение вопроса о том, что именно представляют собой элементы структурных моделей, укажем, что при работе с тезаурусом по ОСД слова и понятия ставятся в связь с двумя рядами фасетов. Одни из них – это обычные общие лексико-семантические категории, как ‘процесс’, ’вещество’ и другие, отнесенность к которым ЛЕ не зависит от синтагматической роли, выполняемой языковой единицей. Другие фасеты имеют характер концептуально-синтагматический и представлены только в контекстах семантических схем индексирования. Решения, связанные с разработкой тезауруса по ОСД, принимаются отдельно для каждой словарной статьи и исходя из того, какой из двух обсуждаемых видов фасетов является более значимым для конечного результата работы системы – обеспечения необходимого уровня точности и полноты результатов автоматизированного поиска информации.

3. На первом этапе работы по фасетной классификации ЛЕ обследуемый материал разносился по 8 семантическим классам, а именно - ДОКУМЕНТЫ, МАТЕРИАЛЫ И ВЕЩЕСТВА, ПРОЦЕССЫ И ОПЕРАЦИИ, ГЕОГРАФИЧЕСКИЕ ПРИВЯЗКИ ОБЪЕКТОВ, ВРЕМЕННЫЕ ПРИВЯЗКИ ОБЪЕКТОВ, ОРГАНИЗАЦИИ, ОБОРУДОВАНИЕ, ДРУГОЙ ЛЕКСИЧЕСКИЙ МАТЕРИАЛ. Далее из этих общих фасетов, как обычно при разработке тезаурусов, выделялись субкатегории (субфасеты) и выявлялись парадигматические связи ЛЕ, в БД ОСД иногда образующие до 6 степеней иерархии логических связей понятий. Так из фасета ПРОЦЕССЫ И ОПЕРАЦИИ был выделен наиболее важный, самый крупный и развернутый в содержательном отношении субфасет ОБЕСПЕЧЕНИЕ СОХРАННОСТИ:

PRESERVATION
.Conservative
Preservation systems
Preserving
Protection
Safeguarding
Безопасность
Защита
Надежность
Обеспечение сохранности
Предохранение
Сохранение
Сохранность
Физическая сохранность
n.ACCIDENT CONTROL
CONSERVATION
CRIME
ENCROACHMENT CONTROL
DISASTER CONTROL
FIRE CONTROL
PRESERVATION EQUIPMENT
SECURITY AND PROTECTION
WARFARE

Связи слов и понятий, входящие в указанный дескрипторный класс, как и других словарных статей ИПТ, выявлялись и устанавливались главным образом на основе того, как эти связи описаны в действующих ГОСТах и специальной литературе по ОСД, то есть на чисто логических основаниях. Предметно-аспектный подход к описанию информации здесь отражается главным образом в том, что большинство относящихся к PRESERVATION ЛЕ представлены лексикой со значением ‘операции’ (DISASTER CONTROL, а не просто DISASTER), а это значение в системе, как выше указывалось, является жестко привязанным к семантической функции ‘аспект’.

Дальнейшая разработка дескриптора PRESERVATION (и других словарных статей ИПТ) заключалась в установлении синонимических связей ЛЕ и логических связей род-вид на нижних уровнях иерархии понятий. В этой работе уже применялась одна из указанных выше пяти предметно-аспектных моделей, а именно – схема анализа информации S [документы]-Attr-P [обеспечение сохранности] – Instr - Loc. С опорой на эту модель информации функции определялось, в частности то, какие лексические единицы являются информативными для БД ОСД и должны быть представлены в ИПТ и какие – не индексируются и не входят в тезаурус. Так, некоторые слова, выступающие

в семантической функции S, (неиндексируемые предметы информации) не представлены в ИПТ, так как эти слова не называют предметов, имеющихся в рассматриваемой схеме анализа информации, а в документе рассматривается, например, “хранение (P), сушка (P) или дезинфекция (P) зерна (S)”, “дизайн (P1) вышивки (P) салфеток (S) и скатертей (S) в Великобритании (Loc) 19 века (Loc)”. Релевантной и ценной для специалистов по обеспечению сохранности документов здесь, очевидно, является не предмет индексирования S, а аспектная часть документов. В некоторых случаях предмет индексирования в документах может отсутствовать, то есть иметь нулевое словесное выражение. Например, в таких документах: “Биологические проблемы (P1) консервации (P) - S (Æ): Специфичность и комплексность”, “Дезинфекция (P) - S (Æ) - в электрическом высокочастотном поле (Instr)“, “Контейнерное хранение (P) - S (Æ) : - Биологические аспекты (P1)”, “Disaster recovery (P) - S (Æ): - Problems and procedures”, “Preventive conservation (P) - S(Æ)”. Отметим еще раз, что в приведенных примерах предмет информации – не операции, а именно нулевое его выражение S(Æ). Такое словесное обозначение предметов не является бессодержательным или “пустым”, но имеет значение ‘любой материальный объект’ – документ, материал, вещество, документохранилище и т.д.

Литература

1. Соколов А.В. Некоторые проблемы типового проектирования информационных тезаурусов// Структурная и прикладная лингвистика: Межвуз. сб. - Вып.1 - 1978. - С.172-180.

2. Оранская Л.И. Некоторые особенности использования дескрипторного поискового языка в библиографической ИПС универсального типа// Науч. и техн. б-ки. - 1997. - №9. - С.13-22.

3. Амхир И. К., Макаров Е.В., Пименов Е.Н., Щербак И.В. Опыт оптимизации фрагмента ИПЯ по технологии огнеупорного производства// Вопросы прикладной лингвистики - Л.: Изд-во ЛГУ, 1978.

4. Пименов Е.Н. Предметно-аспектный подход к анализу и индексированию информации// Предметный поиск в традиционных и нетрадиционных информационно-поисковых системах. - Вып.12 - СПб: РНБ, 1998. - С.96-114.

5. Ranganatan S.R. Colon classification. T.1 Basic classification. - 6th ed. - Bombay: Asia, 1963. - 124 c.

6. Green R. Syntagmatic relationships in index languages// The library quaterly. -Vol.65, N4 - 1995. - Vol.65, N4. - C.365-384.

7. . Fillmore Ch. The case for case// Universals in linguistic theory /Ed. By E.Bach, R.T.Harms - New York; Chicago: Holt, Rinehart and Winston, 1968. - C.1-88.

8. Черный А.И. Введение в теорию информационного поиска. - М.: Наука, 1975. - 238 с.

9. Сахарный Л.В. Предметизация в системе информационного поиска: природа, состояние, проблемы, перспективы // Предметный поиск в традиционных и нетрадиционных информационно-поисковых системах. - Вып.10 - Л.: ГПБ, 1990. - С.14-26.