68th IFLA Council and General Conference
August 18-24, 2002

Code Number: 007-122-R
Division Number: IV
Professional Group: Classification and Indexing
Joint Meeting with: -
Meeting Number: 122
Simultaneous Interpretation: -

Предметный поиск в распределенных ресурсах: краткий обзор новейших тенденций

Мартин Кунц
Die Deutsche Bibliothek
Frankfurt am Main,
Germany

1. Основные положения

Позвольте сначала бросить беглый взгляд в прошлое и вспомнить о том, как выглядел библиотечный мир еще совсем недавно. Если говорить о немецкоговорящем пространстве, то оно характеризовалось, по меньшей мере многие годы, в части предметного раскрытия содержания документов в библиотеках почти неисчерпаемым разнообразием творческих методов и правил как в области классификационного, так и вербального раскрытия содержания. Последнему мы хотим в дальнейшем уделить наше внимание.

Многие из этих методов были абсолютно внутренними решениями, и каждый при этом был, естественно, лучше, чем все другие. Tempi passati. Большое число научных и публичных библиотек общего немецкоговорящего пространства использует для вербального раскрытия содержания документов единый свод правил, а именно Правила предметного каталога (Regeln fuer den Schlagwortkatalog (RSWK)). В соответствии с этим сводом правил, используемым уже свыше 15 лет, создавалась как основа вербального раскрытия содержания документов База данных нормативных предметных рубрик (Schlagwortnormdatei (SWD)). SWD как основа индексирования развилась в универсальный тезаурус немецкоязычного пространства и при этом далеко перешагнула круг своих первоначальных адресатов – университетские и публичные библиотеки. Научные специальные библиотеки также успешно работают с SWD, даже если и не всегда соблюдая RSWK.

Главной целью создания перечня рубрик SWD было раскрытие содержания монографий, которые как и прежде составляют сейчас преобладающую часть публикаций. С другой стороны, различные задачи наших библиотек приводят также к необходимости раскрытия содержания других типов документов. Так, почти незаметно для большей части библиотечной общественности с помощью перечня рубрик SWD уже давно раскрывается содержание журнальных статей, дескрипторы SWD применяются не только для индексирования печатных публикаций, но и других видов информации, таких как музейные объекты или телевизионные передачи. При этом выяснилось, что участвующие в дальнейшем развитии RSWK и SWD были осведомлены о том, что в вопросах терминологии, что означает здесь прежде всего в вопросах пост- и соответственно предкоординации, а также структурированного представления содержания документов, необходимо придерживаться того, что установлено соответствующими нормами в области библиотечного и информационного дела. Благодаря этому возник информационный язык, который по структуре и содержанию соответствует современному уровню развития науки и который тем самым неизбежно нашел заинтересованных в нем и вне библиотечного мира. В принципе перечень рубрик SWD может быть применим для всех типов документов: он, с одной стороны, провинциально тесен, но с другой стороны, по своей универсальности шире профиля отдельной библиотеки.

2. Проблема неоднородности

Сегодняшняя информационная среда делает теоретически возможным осуществление поиска с любого места в любой библиографической базе данных. При этом в базах данных могут быть отражены различные типы документов. Эти документы по мере надобности доступны для предметного поиска с помощью различных методов раскрытия содержания. В сфере вербального раскрытия содержания SWD хотя и имеет широкое обращение, но все же не монополию. Наряду с этим имеют право на существование отраслевые тезаурусы, которые были разработаны в ходе многолетней работы многих информационных центров национального уровня. Даже если бы сложился идеальный случай – применение в заданном языковом пространстве единого универсального тезауруса всеми информационными центрами, – нельзя ожидать, что уважаемые информационные центры заново будут индексировать свои фонды. Мы должны быть рады, что, по крайней мере, библиотеки пользуются одним и тем же информационным языком.

Бросим беглый взгляд на один из таких отраслевых тезаурусов и сравним его с SWD. Как и следовало ожидать, мы находим многочисленные соответствия, но также и некоторые различия. Однако в целом можно сказать, что не имеется практически ни одного понятия, которое не могло бы быть передано с помощью того или иного перечня индексирования. Итак, среди предпочтительных терминов в действующих информационных языках существуют либо конгруэнции, либо эквивалентности 1:1, либо эквивалентности 1:X, последние при этом отражают соответственно И- и ИЛИ-отношения.

То, что в ряде случаев не имеется эквивалентности, не является неожиданным. SWD представляет собой, в сущности, универсальный тезаурус, а от отраслевого тезауруса следует ожидать более специфического списка, даже если принятию этих дескрипторов с помощью SWD при определенных предпосылках, которые можно охарактеризовать как терминологическую совместимость, ничто не мешает. С другой стороны, в проектах, о которых я в дальнейшем хотел бы рассказать, выяснилось, что SWD имеет объемный специфический отраслевой перечень.

Еще недавно соотнести друг с другом такие различные терминологические фонды было фактически невозможно. Сегодня в нашем распоряжении имеются методы, которые могут быть Вам достаточно известны из Интернета, но которые имеют также свои коварства и ограничения. В конце концов все равно, обозначен ли этот метод, к которому стремятся, как перекрестный или как переходный конкорданс. Проще говоря, речь идет о создании связей между эквивалентными терминами обоих тезаурусов, представляющими одинаковые понятия, т.е. иными словами, речь идет о совместимости информационных языков. При этом есть надобность в некой связывающей

тезаурусы системе, которая управляет имеющимися связями, изменениями в них, введением новых связей и которая одновременно помогает пользователям при навигации между современными фондами данных.

3. Проекты

Дальше я хотел бы набросать эскизы трех проектов, которые похожи по своей структуре. Два из них являются одноязычными (CARMEN, Перекрестный конкорданс по экономике), третий – многоязычным (MACS).

3.1. CARMEN

Часть проекта CARMEN посвящена сопряжению тезауруса Информационного центра по социальным наукам (IZT) с SWD. Метод, посредством которого определяются эквиваленты и устанавливаются связи, заманчиво прост: исходя из алфавитных перечней, которые содержат предметные рубрики из специфической предметной области, интеллектуально определяются связи между обоими тезаурусами. Не так давно подобный метод действий показался бы, пожалуй, устаревшим, и каждый, разумеется, счел бы необходимым сделать поисковую машину лучше, чем все предыдущие. Но безграничная эйфория о мощности таких продуктов, кажется, уже прошла, прошла также фаза, когда об одном ялике говорилось десять: "позвольте нам идти вперед и строить поисковую машину, чтобы мы сделали себе имя". "Широко распространенная точка зрения, что бурный прогресс техники одарил бы нас уже в ближайшем будущем и одним махом сделал бы ненужными и нелюбимый терминологический контроль и стандартизацию, относится к области магии или к языку рынка" (Christof Wolters. Ist die Schlagwortnormdatei fuer die Objektdokumentation im Museum geeignet? – AKMB-News 1/1998).

Мы отбросили также в рамках данного проекта идею автоматического ранжирования с помощью метода, который еще только разрабатывается. Такой метод мог бы обеспечивать, помимо машинной побуквенной идентификации дескриптора, его интеллектуальную проверку: подходит ли указанным буквам данное понятийное содержание. Вместо этого мы бодро принялись за интеллектуальную работу и в первом полугодии с одной научной сотрудницей, работающей на полставке, так обработали 3.500 предметных рубрик, что эквивалентные соотношения между тезаурусом Информационного центра по социальным наукам, тезаурусом Немецкого института педагогических исследований и SWD были определены, и они могут быть включены в систему управления связями. Установлено, что преобладают конгруэнции. Так как мы здесь находимся в одноязычной среде, это, что называется, в природе вещей. Хотя и даны альтернативы различных дескрипторов для одинаковых понятий, специалисты выбирают, однако, большей частью те же самые часто употребляемые термины. В целом при сравнении тезаурусов обнаружено около 85% эквивалентов, из них 75% конгруэнтных. При этом выяснилось, что SWD имеет в распоряжении явно более обширный отраслевой перечень, чем оба отраслевых терминологических фонда.

3.2. Перекрестный конкорданс по экономике

Как и в проекте CARMEN, здесь также идет речь об установлении связей между признаваемыми эквивалентными дескрипторами стандартного тезауруса по экономике и SWD. На этот раз мы будем автоматически проводить первый контроль на дублеты, иначе, конечно, чем в CARMEN. Мы не знаем, будет ли этот метод иметь смысл, сэкономит ли нам это также время на интеллектуальную работу или приведет, возможно, к большей путанице. В любом случае нам важно получить опыт для дальнейших проектов, чтобы потом мы больше не руководствовались субъективными теориями.

3.3. MACS

1997 г. был отмечен переломом и качественным прогрессом в более чем десятилетнем существовании SWD. В последние годы снова и снова требовалось включение в SWD иноязычных терминов. В областях, в которых научный язык и язык публикаций в немецкоязычном пространстве является в основном английским, в SWD всегда включаются англоязычные ссылки. Англоязычные ссылки действуют также с итальянскими и французскими терминами по культуре, применяемыми в отраслевой терминологии. Приписка таких ссылок ориентирована при этом в целом на потребности немецкоязычной SWD. До недавних пор не были учтены требования иноязычной литературы и международного обмена данными по предметному раскрытию содержания документов в глобальной библиотечной сети. Первые попытки Отдела предметного раскрытия содержания документов Союза немецких переводчиков закрыть эти информационные лакуны относятся к началу последнего десятилетия. Благодаря недавней инициативе Швейцарской государственной библиотеки смог стартовать проект MACS (Многоязычный доступ к предметным заголовкам). В этом проекте наряду со Швейцарской государственной библиотекой приняли участие Национальная библиотека Франции, Британская библиотека и Немецкая библиотека. Целью MACS является разработка связей между тремя объемными базами данных предметных рубрик – Библиотеки Конгресса, Национальной библиотеки Франции (база данных RAMEAU) и SWD. Ближайшая цель проекта состоит в том, чтобы в каждой нормативной базе в выбранных отраслевых областях обозначить соответствующие эквивалентные предпочтительные термины другой нормативной базы. В конце концов работа ограничилась темами Спорт и Театр, а также выбором особенно часто используемых дескрипторов. Эквиваленты между дескрипторами были разработаны интеллектуально без поддержки автоматизированных систем, также как в проекте CARMEN. Мы видели свою задачу не в том, чтобы привести проблему автоматического перевода к ее окончательному решению. Мы стремились при этом развить программу, которая по единственному слову распознает его однозначное соответствие в другом языке. "Машинный перевод существует? Честный ответ на этот вопрос – НЕТ" (S. Krauwer. Machine Translation: State of the Art, Trends and User Perspective. – TELRI-Proceedings of the First European Seminar "Language Resources for Language Technology". Budapest, 1996).

Создание полного многоязычного тезауруса не выдвигается в качестве задачи. Этот тезаурус потребовал бы соблюдения свойства изоморфизма. Это означает, что должны существовать эквиваленты не только в части предпочтительных терминов, но и в части структурирования, что каждому понятию в каждом сопряженном языке на той же иерархической ступени должен быть поставлен в соответствие эквивалент. К тому же требуется приспособить структуры национальных нормативных баз данных друг к другу и в данном случае закрыть имеющиеся лакуны. Полный конечный продукт мог бы в конце концов заменить национальные нормативные базы данных, так они были бы интегрированной частью фонда MACS. Эта классическая подготовка многоязычного тезауруса не имеет шансов быть когда-либо реализованной. В результате практических опытов проекта MACS и ознакомления с существующими многоязычными отраслевыми тезаурусами появился большой скепсис: имеют ли еще смысл многоязычные тезаурусы в чистой форме при наличии соответствующих национальных и международных норм, а также, вообще, оправдывает ли значительные затраты разработка подобных средств индексирования.

Основной целью проекта MACS является, конечно, охват всех отраслевых областей. Открытие базы данных для других языков было бы следующим логическим шагом. Метод, к которому стремится MACS, не будет касаться структур теперешних национальных нормативных баз данных, таких как CARMEN и тезаурус IZT. Это не исключает, что в ходе работ мы не подойдем к изменениям теперешних тезаурусов, но при этом речь идет о корректурах или уточнениях, которые раньше или позже все равно были бы сделаны и, как побочный продукт, повлекли бы за собой качественное улучшение отдельных баз данных.

Пользователь осуществляет поиск на своем родном языке или в хорошо знакомой ему языковой среде и может потом работать с указанными эквивалентами в подобных аффилированных библиографических базах данных. Необходимую технологию заново развивать не нужно, мы можем довериться Интернету, отдельные протоколы имеются (Z39.50), и все это может применяться для связи нормативных баз данных с базами данных заглавий – они доступны в сети.

4. Пользователи

Безразлично, какие виды документов отражены в базе данных, традиционные или электронные публикации, идет ли речь о фильмах, музейных объектах или архивах, имеются ли классические описания под заглавием или описания в формате Дублинского ядра, полнотекстовые версии или цитаты, – проблема пользователя остается неизменной. Как и прежде друг другу противостоят две системы, в которых различным образом организовано знание. В качестве вербальной связи между пользователями, с одной стороны, и библиографическими базами данных, с другой стороны, служит тезаурус, который наряду с упорядочиванием внутри системы базы данных выполняет двоякие функции:
1) поддержка запроса пользователей;
2) структурированная навигация в базах данных.

MACS интеллектуально связывает определенными эквивалентами содержание библиографических баз данных, в которых для раскрытия содержания применяются терминологически контролируемые перечни и которые представляют также упорядоченный структурированный фонд данных. Мне кажется поэтому мало ценным, иначе регрессом, если при подобных фондах данных – это было бы при наших намерениях благодарной альтернативой – применяется автоматический метод с электронными словарями. Для MACS это означало бы, что имеющийся в современных системах однозначный понятийный порядок был бы нарушен и пользователь мог бы осуществлять навигацию от многозначности до многозначности.

Если пользователь отлично знает свою иноязычную отраслевую терминологию, он не нуждается в таком продукте, а если он знает ее плохо, тогда это ему не поможет. Однако связь от предпочитаемого термина в одном тезаурусе к его эквиваленту в другом могла бы помочь пользователю. По команде Просмотр пользователь получает список возможных дескрипторов на других языках. Этот список – автоматически производимый продукт, в котором связи терминов определены интеллектуально. Список показывает проблемы, которые должна поставить перед собой Система управления международным тезаурусом, а именно, разработку эквивалентов 1:X, которые возникают вследствие различных терминологических традиций информационных языков. Чем больше терминология основывается на лексикографической базе, тем проще связь. Так как Перечни предметных рубрик Библиотеки Конгресса, а также, но в меньшей степени, RAMEAU являются предкоординированными перечнями, часто требуется больше дескрипторов SWD, чтобы с помощью И- или ИЛИ-логики, являющейся средством сближения неточных эквивалентов, получить более точные эквиваленты. Если число задействованных языков возрастает, тогда становится, естественно, более комплексным управление связями, – это относится также к привязке немецкоязычного отраслевого тезауруса к SWD. Чтобы не стать здесь жертвой комбинаторного взрыва, стоило бы упомянуть, что среди новых информационных языков останавливают свой выбор только на одном языке, преимущественно на том, который относительно признака посткоординации - предкоординации наиболее похож на естественный язык.

MACS основывается на том, что пользователь сам с помощью тезаурусов добивается результатов интеллектуального вербального раскрытия содержания. Мы должны это делать с библиографическими фондами данных значительного объема. Простые иерархии, привычные Вам, не справятся с этими объемами, в то время как тезаурусы являются более сильными и комплексными языками индексирования, позволяющими пользоваться синтаксическими методами

индексирования. Если в проектах связи неоднородных фондов число документов возрастет, тогда сразу должно стать благоразумной необходимостью возможное как раз благодаря нашему методу индексирования и уже имеющееся в наличии предложение по структурируемой среде пользователей.

5. Краткое резюме

Три проекта показывают, что возможно установление связей между неоднородными фондами данных при обозримых интеллектуальных и не слишком больших технических издержках. Мы видели, как фонды, содержание которых отражено с помощью различных информационных языков, могут быть связаны друг с другом. Предпосылкой для такого метода должно быть то, что речь идет об объемных и известных фондах с информационными языками, являющимися полностью продуктом упорного труда. Целью нашего метода не может быть поддержка частных терминологических ресурсов небольших организаций или содействие разработке параллельных ресурсов.