Подходы к описанию и использованию тезаурусов в информационных системах

© Аджиев Алим Сапарович
ВЦ РАН
ajiev@ccas.ru

© Нгуен Хунг Мань
ВЦ РАН
nmhungru@yahoo.com

Аннотация

В статье рассмотрены разные подходы к формализации тезаурусов, а также стандарты ISO и ANSI. Сделан анализ некоторых возможных платформ для такой формализации, описаны особенности работы с тезаурусами в информационных системах, а также проблемы при этом возникающие, требования к реализации тезауруса в рамках Semantic Web [12]. Рассмотрены особенности и различия классификаторов ресурсов и обычных терминологических и лингвистических тезаурусов. Дан сравнительный анализ существующих схем данных и подходов к реализации тезаурусов для информационных систем на основе RDF. Рассмотрены также вопросы организации пользовательских интерфейсов для работы с тезаурусами, и использования их при поиске в информационной системе, а также интерфейсы администрирования тезаурусов. Во второй части статьи на основании проделанного анализа сформулированы требования к описанию тезауруса в ИСИР [17], и приведена схема данных для представления тезауруса в этой информационной системе, удовлетворяющая перечисленным требованиям, и небольшой пример реализации в ней классификатора MSC.

Тезаурусы в описании информации

Для описания какой-либо предметной области всегда используется определенный набор терминов, каждый из которых обозначает или описывает какое-либо понятие или концепцию из данной предметной области. Совокупность терминов, описывающих данную предметную область, с указанием семантических отношений (связей) между ними является *тезаурусом*. Такие отношения в тезаурусе всегда указывают на наличие смысловой (семантической) связи между терминами.

Основным отношением (связью) между терминами в тезаурусе является связь между *более широкими* (более выразительными) и *более узкими* (более специализированными) понятиями. Часто выделяют 3 подвида этого отношения:

- Один термин обозначает понятие, являющееся частью понятия, обозначаемого другим термином (например, «наука» и «математика», «математика» и «теория чисел»).
- Один термин обозначает частный случай понятия, обозначаемого другим термином (например, «птицы» и «попугаи»)
- Один термин обозначает элемент класса, обозначаемого другим термином («горные районы» и «Кавказ»).

Это отношение на множестве терминов является отношением частичного порядка, то есть множество терминов с такими связями образует ациклический граф, или полииерархическую структуру.

Существуют также и другие связи между терминами. Например, одно понятие или концепция может быть обозначено несколькими терминами, являющимися синонимами. Некоторые термины могут быть антонимами для других. Часто среди терминов, относящихся к одному понятию, выделяют единственный (для каждого языка тезауруса) *наиболее предпочтительный* (*наиболее подходящий*) *термин*, который наиболее хорошо характеризует, или обозначает данное понятие. Остальные термины являются *менее предпочтительными* (*менее подходящими*).

Помимо вышеописанных, между терминами могут существовать также и другие, *ассоциативные связи*, если понятия, обозначаемые этими терминами, как-либо связаны между собою по своему смыслу, за исключением описанных выше иерархических связей.

В многоязычных тезаурусах существуют также *связи эквивалентности* между терминами на разных языках. Выделяют полную (строгую) эквивалентность, и несколько видов частичной (нестрогой) смысловой эквивалентности терминов на разных языках.

Тезаурус часто содержит *комментарии* к терминам, раскрывающие для пользователя смысл термина, а также поясняющие, как следует его использовать.

Тезаурусы применяются, прежде всего, для классификации и поиска информационных ресурсов. При этом каждому ресурсу могут быть сопоставлены одно или более понятий, описываемых терминами в тезаурусе, а пользователь, осуществ-

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

ляющий поиск, может по тезаурусу найти интересующие его понятия в данной предметной области, а также все характеризующие их термины. То есть на основе связей тезауруса происходит расширение поискового запроса (расширение слов запроса синонимичными, более общими или более частными по смыслу терминами). Навигация по связям тезауруса помогает четче сформулировать сам запрос.

Существует ряд стандартов разного уровня значимости и проработанности на формат представления тезаурусов. Эти стандарты представляют тезаурус в виде набора объектов нескольких типов, между которыми может быть несколько типов связей. Некоторые стандарты (например, стандарт ANSI/NISO Z39.19-1993) регламентируют также формат представления тезауруса в линеаризованном (текстовом) виде, пригодном для восприятия, как машиной, так и человеком.

Стандарты ISO и ANSI/NISO Z39.19-1993

Основными документами, регламентирующим формат представления тезауруса, являются стандарты ISO 2788-1986 для описания одноязычных тезаурусов, и ISO 5964-1985 для многоязычных.

Стандарт ISO 2788-1986 определяет тезаурус, как набор терминов, связанных между собою соответствующими связями (отношениями).

Термины могут иметь следующие атрибуты:

- ***SN*** **–** *Scope Note.* Комментарий к термину. Например, представляет вербальное пояснение термина, или правила его использования.
- ***TT*** **–** *Top Term.* Признак, Выделяющий термины на самом верхнем уровне иерархии (термины наиболее общих понятий в данной иерархии понятий).

Связи между терминами могут быть следующими:

- ***USE*** **–** Связывает термин с наиболее предпочтительным (на том же языке) термином для данного понятия. *A* ***USE*** *B* = термин *B* является наиболее предпочтительным для понятия, обозначаемого термином *A*.
- ***UF*** **–** *Used For.* Обращение связи ***USE***. Связывает наиболее подходящий термин с синонимами и квазисинонимами (менее подходящими терминами). *A* ***UF*** *B* ⇔ *B* ***USE*** *A*.
- ***BT*** **–** *Broader Term.* Связь термина с термином более общего понятия. *A* ***BT*** *B* = термин *B* обозначает более общее понятие по сравнению с понятием, обозначаемым термином *A*.
- ***BTG*** **–** *Broader Term Generic.* Вариант связи ***BT*** в случае, когда термин характеризует разновидность понятия, определяемого более общим термином. Например, «Попугаи» и «птицы». Наличие связи ***BTG*** подразумевает наличие связи ***BT***.
  A ***BTG*** *B* ⇨ *A* ***BT*** *B*.
- ***BTP*** **–** *Broader Term Partitive.* Вариант связи ***BT*** в случае, когда термин характеризует часть понятия, определяемого более общим термином. Например, «математика» и «теория чисел». Наличие связи ***BTP*** подразумевает наличие связи ***BT***.
  A ***BTP*** *B* ⇨ *A* ***BT*** *B*.
- ***NT, NTG, NTP*** **–** *Narrower Term, Narrower Term Generic, Narrower Term Partitive.* Обращение связей ***BT, BTG*** и ***BTP*** соответственно.
  A ***NT*** *B* ⇔ *B* ***BT*** *A; A* ***NTG*** *B* ⇔ *B* ***BTG*** *A; A* ***NTP*** *B* ⇔ *B* ***BTP*** *A*.
- ***RT*** **–** *Related Term.* Ассоциативная связь. Связывает семантически связанные между собою термины, не находящиеся при этом в одной иерархии, и не являющиеся синонимами или квазисинонимами. Эта связь проставляется в тех случаях, когда пользователю тезауруса может быть полезно осуществлять поиск или индексацию не только по данному термину, но и по связанному с ним. Связь должна быть двунаправленной (симметричной):
  A ***RT*** *B* ⇔ *B* ***RT*** *A*.

Структура многоязычных тезаурусов регламентируется стандартом ISO 5964-1985. В нем, помимо всех вышеперечисленных связей и требований к ним, есть также связи между эквивалентными терминами на разных языках. Существуют следующие типы таких связей:

- *Полная эквивалентность*
- *Неполная эквивалентность* (значения терминов не совпадают, но пересекаются)
- *Частичная эквивалентность* (значение одного термина шире, чем значение другого)
- *Эквивалентность один ко многим* (значение одного термина соответствует совокупности значений нескольких терминов).

Американский стандарт ANSI/NISO Z39.19-1993 расширяет и уточняет стандарт ISO 2788-1986 для одноязычных тезаурусов, а также накладывает ряд дополнительных ограничений на структуру тезауруса. Основные его отличия следующие:

Добавлены новые связи между терминами:

- ***BTI*** **–** *Broader Term Instance.* Вариант связи ***BT*** в случае, когда термин характеризует элемент класса, или частный случай понятия, определяемого более общим термином. Например, «Кавказ» и «горные районы». Наличие связи ***BTI*** подразумевает наличие связи ***BT***.
  A ***BTI*** *B* ⇨ *A* ***BT*** *B*.
- ***NTI*** **–** *Narrower Term Instance.* Обращение связи ***BTI.***
  A ***NTI*** *B* ⇔ *B* ***BTI*** *A*.
- ***GS*** **–** *Generic Structure.* Это иерархическая связь, используемая для визуального представления тезауруса. Она может не соответствовать структуре связей ***BT/NT***. Эта связь используется потому, что визуальное представление полииерархической структуры, образуемой связями ***BT/NT*** затруднительно и ненаглядно.
- ***USE+*** **–** *Use ... and…* Связь один ко многим. Используется, когда для данного термина более предпочтительными является совокупность не-

скольких терминов. Например, «Угольные шахты» ***USE+*** «Уголь» and «Шахты».

- ***UF+*** – Обращение связи ***USE+***.

Добавлены также атрибуты термина:

- ***ID*** – *Identifier.* Уникальный идентификатор термина.
- ***HN*** – *History Note.* История модификации связей и атрибутов данного термина.

В стандарте указаны следующие ограничения на структуру тезауруса:

- Из термина, не являющегося наиболее подходящим для какой-либо концепции, могут исходить только связи ***USE*** и ***USE+***, а входить только связи ***UF*** и ***UF+***. Никаких других связей этот термин иметь не может.
- Термин не может иметь связи с самим собою.
- Одна пара терминов не может иметь 2 или более связи (за исключением случаев, когда одна связь следует из другой по правилам стандарта).

Стандарт ANSI/NISO Z39.19-1993 помимо структуры регламентирует также и другие аспекты создания, представления и поддержки тезаурусов. Однако это выходит за рамки рассмотрения данной статьи.

Следует заметить, что данный стандарт не может полноценно представлять в тезаурусе антонимы. Наиболее предпочтительный термин, имеющий антоним, должен быть связан с ним связью *Use*. При этом термин-антоним не может иметь других связей, кроме этой. Однако иногда антоним сам по себе реализует одно из представленных в тезаурусе понятий, включенное, в частности, в общую иерархию понятий, и является для него наиболее предпочтительным термином. Например, в вычислительной математике есть «линейные вычислительные задачи» и «нелинейные вычислительные задачи». В этом случае в рамках данного стандарта между ними нельзя установить отношение, показывающее антонимичность этих понятий.

Особенности применения тезаурусов в информационных системах.

Модель данных

Описанные выше стандарты были разработаны для представления тезаурусов в виде, удобном для ручной индексации информационных ресурсов. Такая модель может быть также использована для машинной индексации с целью осуществления последующего поиска по ключевым словам.

Однако существует ряд тезаурусов, основная задача которых не индексация ресурсов, а их классификация. В этом случае основными объектами таких тезаурусов (*классификаторов*) выступают не термины, а понятия (*рубрики*), и, часто, идентифицирующие их уникальные идентификаторы (коды классификации). Отношения в таком тезаурусе – не семантические связи между терминами, а характеризующие логику описываемой предметной области отношения между понятиями (рубриками). Примерами таких тезаурусов могут служить тематические классификаторы в разных отраслях науки, например, MSC [13], PACS [14], DDC [15].

Структура классификатора соответствует структуре обычного тезауруса, поскольку связи между его рубриками по смыслу те же, что и между терминами тезауруса, и классификатор является его частным случаем. Однако при классификации в соответствие ресурсам ставятся не термины, а обозначаемые ими понятия. Потому в схеме данных информационной системы понятия тезауруса должны быть выделены в самостоятельные объекты. Это означает, что такая схема должна иметь структуру, отличную от вышеописанных стандартов, в которых понятия не выступают отдельными объектами, а есть лишь термины и связи между ними. В то же время, схема должна позволять работать с тезаурусами, описанными в соответствии с этими стандартами, т.е. быть совместима с ними.

Среди связей между терминами в вышеописанных стандартах следует различать связи, которые по смыслу характеризуют фактически соотношения не между терминами, а между термином, и обозначаемым им понятием. К таковым относятся связи *Use* и *Used For*. В схеме данных для информационной системы стоит ставить такие связи между понятиями и терминами, которые их обозначают.

Аналогично, иерархические и ассоциативные связи по смыслу являются связями между понятиями. Признак *Top Term* также является признаком понятия, находящегося на вершине иерархии понятий.

Таким образом, получается следующее отображение связей между терминами в стандартах ISO и ANSI для одноязычных тезаурусов на отношения в схеме данных информационной системы: Те связи, которые допустимы между наиболее предпочтительными терминами для каких либо понятий, в схеме данных информационной системы становятся отношениями между понятиями. Те связи, которые были допустимы между наиболее предпочтительным термином и другими терминами данного понятия, становятся отношениями между понятием и термином.

Как указывалось выше, в многоязычных тезаурусах термины имеют атрибут *язык*, на котором данный термин обозначает данное понятие. Кроме того, стандартом ISO 5964-1985 предусматривается ряд отношений эквивалентности между терминами на разных языках, допускающие, помимо строгой эквивалентности, несколько видов неполной эквивалентности терминов. По смыслу атрибут *язык* – свойство термина, а не понятия. В то же время термины на разных языках, между которыми есть только частичная эквивалентность, строго говоря, соответствуют разным, пусть и близким, понятиям [6].

Таким образом, более естественной в схеме данных тезауруса для информационных систем будет привязка языка к терминам, а не к понятиям. Более того, такой подход является единственно возможным для классификаторов, в которых именно независящие от языка понятия классифицируют другие

ресурсы. Обычно такие классификаторы изначально создаются как одноязычные, и лишь потом для них делаются переводы на другие языки. В этом случае между терминами на разных языках имеет место только строгая эквивалентность, поскольку при переводе для каждого термина дается его строгий эквивалент (который является эквивалентом по определению, в контексте данного классификатора, даже если фактически перевод не совсем точен). Привязка языка к понятию означала бы необходимость делать отдельную копию одного и того же понятия для каждого языка, и делать отдельную связь между каждой копией понятия и классифицируемым им ресурсом. Привязка языка к термину привязать все эквивалентные термины на разных языках к одному и тому же понятию.

Однако в тезаурусах, где много отношений неполной эквивалентности между разноязычными терминами, а также имеются разные иерархии для терминов на разных языках, даже полностью эквивалентные термины могут оказаться в разных иерархиях, а значит, не могут быть привязаны к одному понятию. Все это означает, что для поддержки многоязычных тезаурусов схема данных должна предусматривать описанные в стандарте ISO соотношения эквивалентности между терминами на разных языках, как отношения между понятиями. При этом для каждого тезауруса, в зависимости от его специфики, необходимо сделать выбор, каким образом реализовывать отношение полной эквивалентности между разными терминами:

1. Приписывать термины к разным понятиям, и ставить между понятиями отношение полной эквивалентности.
2. Приписывать термины к одному и тому же понятию.

Очевидно, для классификаторов необходимо использовать второй подход, а для многоязычных тезаурусов, имеющих разные иерархии на разных языках – первый. Следует заметить, что тезаурус, в котором есть отношение неполной эквивалентности, по смыслу уже подразумевает наличие разных иерархий на разных языках, а значит, необходим первый подход при их реализации.

Еще одним важным атрибутом термина в тезаурусе является комментарий к нему (*Scope Note*). В тезаурусах-классификаторах, где, по сути, первично понятие, а не термин, комментарий, как правило, также характеризует понятие. Однако, в других тезаурусах комментарий может относиться именно к термину. Например, описывать случаи предпочтительного употребления именно этого синонима перед другими. Таким образом, в разных тезаурусах комментарии могут относиться, как к понятиям, так и к терминам. Выбор зависит от конкретного тезауруса. Универсальная схема данных в информационной системе должна допускать оба варианта применения комментариев.

Платформа реализации тезауруса, требования Semantic Web

Модель данных тезауруса, в том числе и учитывающая все перечисленные выше требования, может быть создана практически на любой платформе представления онтологий. В частности, существуют модели тезаурусов на основе Topic Maps [11], RDF [4, 6], DAML [5].

Однако для того, чтобы реализация тезауруса могла в полной мере соответствовать концепциям проекта Semantic Web, на нее накладываются следующие требования:

1. *Синтаксическая и семантическая интероперабельность.* Любое приложение, работающее в соответствии с требованиями Semantic Web должно иметь возможность работать с тезаурусом без предварительного согласования форматов.
2. *Расширяемость тезаурусов.* При необходимости любое приложение должно иметь возможность может добавить в открытый тезаурус свои элементы, и использовать его в таком расширенном виде для своих нужд.
3. *Расширяемость модели.* Схема данных должна допускать расширения и детализацию. То есть любое приложение должно иметь возможность добавить в модель новые типы ресурсов и связей, в частности детализировать уже существующие, если это, например, необходимо для описания нестандартного тезауруса. В то же время приложения, не знающие о таком расширении, должны иметь возможность корректно работать с этим тезаурусом в рамках прежней модели, имея доступ к той части данных тезауруса, которая в нее вписывается.

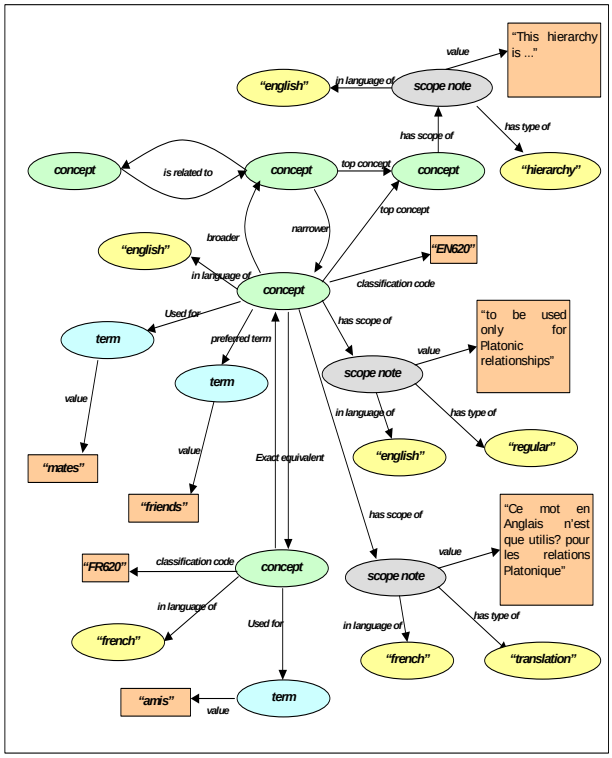
Эти требования накладывают ограничения и на платформы реализации тезауруса. Например, платформа Topic Maps [10] в формате XTM в целом удовлетворяет пунктам 1 и 2, но не удовлетворяет пункту 3. Наиболее соответствует перечисленным требованиям платформа RDF, а так же ее расширения (например, DAML+OIL) [19]. Платформа RDF принята также в качестве основной для описания онтологий в Semantic Web.

Подходы к описаниям тезаурусов

В этом разделе рассмотрены некоторые существующие схемы данных на основе RDF, предложенные в качестве стандартов для описания тезаурусов в информационных системах.

Формат представления многоязычного тезауруса в RDF, разработанный в рамках проекта LIMBER.

Данный формат изначально разрабатывался для многоязычного тезауруса ELSST (European Language Social Science Thesaurus) [4]. Однако в настоящий момент LIMBER [3] предлагает данную модель как универсальную, для представления многоязычных тезаурусов.



Пример описания тезауруса в схеме данных LIMBER

Модель имеет следующие основные типы объектов (ресурсов):

- *Понятие (Consept)*
- *Термин (Term)*
- *Комментарий (ScopeNote)*
- *Язык (LanguageCode)*

Существуют следующие свойства понятий:

- *Уникальный идентификатор (ClassificationCode)*
- *Язык (inLanguageOf)*
- *Комментарий к понятию (hasScopeNote)*
- *Наиболее предпочтительный термин (PreferredTerm)*
- *Менее предпочтительный термин (UsedFor)*

Существуют следующие свойства комментариев:

- *Язык (inLanguageOf)*
- *Тип (hasTypeOf)*. Существуют следующие типы комментариев:
  - *General*. Комментарий к понятию на основном языке тезауруса (один из языков тезауруса в модели выделяется как основной или главный).
  - *Translation*. Комментарий на неосновных языках.
  - *Hierarchy*. Признак понятия, находящегося на вершине иерархии.
  - *History*. Пометки об истории изменения этого понятия в предыдущих версиях тезауруса.

Существуют следующие связи между понятиями одного языка:

- *Более широкое понятие (BroaderConsept)*
- *Более узкое понятие (NarrowerConsept)*
- *Связанное понятие (isRelatedTo)*
- *Указатель на корневую концепцию данной иерархии (TopOfHierarchy)*

Существуют следующие связи между понятиями на разных языках:

- *Строгая эквивалентность (ExactEquivalent)*
- *Нестрогая эквивалентность (InexactEquivalent)*
- *Частичная эквивалентность (PartialEquivalent)*
- *Эквивалентность типа «один ко многим» (OneToManyEquivalent)*

Эта модель хорошо подходит для описания многоязычных тезаурусов, в которых существуют разные иерархии терминов на разных языках. Однако здесь язык термина является атрибутом понятия, а не термина. Как было описано выше, такая модель неудобна для описания многоязычных классификаторов ресурсов, в которых понятия семантически не связаны с каким-либо определенным языком.

Схема данных тезауруса ILRT

Эта схема данных строилась в расчете на работу не только с тезаурусами в обычном, «лингвистическом» смысле, но и с классификаторами. Потому язык термина привязан не к понятию, а к самому термину, а термины на разных языках, точно эквивалентные друг другу, привязаны к одному и тому же понятию. Термины на разных языках, не имеющие строгой эквивалентности, должны быть отнесены к разным понятиям.

Модель предполагает 2 уровня детализации описания тезауруса. Первый уровень реализует связи, предусмотренные стандартом ISO 2788-1986 для одноязычных тезаурусов, а также атрибут «язык» для терминов. Второй уровень детализации пока не оформлен в виде RDFS, и предполагает детализацию ряда связей 1 уровня детализации. Например, связь «более общее понятие» распадается на 3 RDF-связи, реализующие 3 описанных выше вида этой связи. Аналогично происходит детализация других связей.

По сути, эта схема предназначена для одноязычных тезаурусов и для тезаурусов-классификаторов, поскольку механизм полной поддержки многоязычных тезаурусов никак не прописан, а обозначено только направление, как это можно сделать в рамках данной модели.

Особенностью данной модели, в сравнении с предыдущей, является отсутствие избыточных связей оптимизирующих скорость исполнения запросов. Например, нет связи «более широкое понятие», поскольку оно является обращением связи «более узкое понятие». Отсутствует также связь понятий с самыми верхними понятиями включающих их иерархий, поскольку она тоже вычисляется из иерархических связей. Это накладывает дополнительные ограничения на техническую реализацию такой модели. В частности, традиционные способы реализации графов не позволят за один шаг вычислить корневую вершину иерархии для произвольного понятия.

Модель тезауруса DRC

Эта модель наиболее точно соответствует модели одноязычного тезауруса ISO 2788-1986. В частно-

сти, в нем отсутствует класс понятий, и все связи существуют только между терминами. Некоторые связи детализированы, в частности выделены разные виды связей менее предпочтительными терминами. Модель реализована на языке DAML [16].

Стоит выделить одну явную ошибку этой модели. Связь *Related Term* является транзитивной, что не соответствует действительности. Например, связанными терминами являются *транспортировка нефти* и *трубы для нефтепроводов*, а также *трубы для нефтепроводов* и *стальной прокат*. Однако прямой связи между понятиями *транспортировка нефти* и *стальной прокат*, очевидно, нет [4].

Поскольку в модели нет понятий, как отдельных объектов, она не удобна для реализации классификаторов.

Интерфейсы работы с тезаурусом в информационных системах

Просмотр тезауруса и поиск ресурсов

В информационной системе тезаурус является не только самостоятельным информационным ресурсом, но и инструментом для классификации или индексации ресурсов. Таким образом, пользователь информационной системы должен иметь возможность:

- Осуществлять просмотр тезауруса.
- Осуществлять поиск ресурсов по ассоциированным с ними терминам или понятиям.

Поиск ресурсов может вестись двумя способами:

- Поиск по ключевым словам, используя тезаурус.
- Навигация по тезаурусу. То есть поиск сначала нужного понятия в тезаурусе с последующим запросом ресурсов, соответствующих этому понятию.

При поиске ресурсов по ключевым словам поисковая система может, используя тезаурус, расширять результаты поиска, выдавая пользователю не только ресурсы, соответствующие введенным пользователем ключевым словам, но и ресурсы, соответствующие связанным с ними терминам, или терминам, обозначающим также более узкие понятия относительно исходного термина. Например, если пользователь ищет ресурсы, соответствующие термину «туннель», в результатах поиска необходимо выдать также все ресурсы, соответствующие термину «тоннель», поскольку оба они являются разными вариантами написания одного и того же слова. Или если ищутся ресурсы, соответствующие понятию *дифференциальные и функциональные уравнения*, имеет смысл включить в результаты поиска также ресурсы, соответствующие рубрике *системы функциональных уравнений и неравенства*. Система поиска может также, используя тезаурус, подсказать пользователю, по каким еще словам ему стоит осуществить поиск (например, квазисинонимы, связанные термины, более широкие термины, и т.д.). Оба этих варианта использования тезауруса широко применяются, например, в поисковых машинах Internet.

Интерфейс просмотра тезауруса должен:

- Показывать все атрибуты данного термина или понятия.
- Показывать, с какими терминами и понятиями связан данный термин или понятие.
- Достаточно наглядно показывать пользователю место термина или понятия в иерархии понятий тезауруса.

Первые 2 пункта выполнимы, если показывать пользователю для каждого понятия тезауруса на отдельном экране (странице) все его атрибуты, все связанные с ним термины (на всех или на определенном языке), и все связанные с ним понятия. Интерфейс должен при этом обеспечивать переход к странице просмотра любого из перечисленных на данной странице понятий. Если в тезаурусе схемой данных разрешена привязка термина более чем к одному понятию, на той же странице для каждого термина должны быть перечислены также понятия, к которым еще привязан данный термин. Если у понятия есть термины на других языках, не полностью эквивалентные данному понятию, или полностью эквивалентные, но прикрепленные в силу структуры данного тезауруса к другим понятиям, на странице должны присутствовать ссылки на страницы этих понятий.

Наглядно показать пользователю место термина или понятия в тезаурусе достаточно сложно, поскольку достаточно наглядное отображение полииерархической структуры на одной странице, в отличие от иерархии, сложно, как для отображения, так и для восприятия пользователем. В общем случае невозможно будет обойтись без пересекающихся линий, показывающих иерархические связи между понятиями. Потому будет правильно показать только часть понятий и связей, которая, с одной стороны, была бы легко отображаемой и воспринимаемой, и в то же время достаточно наглядно показывала бы место понятия в общей иерархии понятий.

Если тезаурус имеет строго древовидную структуру, то представление дерева обычно осуществляется следующими способами:

1. Отрисовка пути по дереву от корня к текущему элементу. Например, [18].
2. Отрисовка пути по дереву от корня к текущему элементу, а также соседей каждого предка текущего элемента. Например, [17]
3. Отрисовка всего дерева целиком. Обычно в таких случаях пользователь может открывать и закрывать отображение на экране потомков любых узлов. Например, программа «Проводник» («Explorer») в операционных системах Microsoft Windows.

При обычной реализации дерева в реляционной БД, когда все элементы дерева хранятся в одной таблице, и каждый элемент имеет атрибут-указатель на непосредственного предка, для отрисовки дерева в первых двух случаях можно обойтись одним запросом к БД. Это достигается хранением избыточной информации в таблице дерева.

В случае полииерархической структуры первые 2 из вышеописанных способов также могут быть применены. Но в этом случае необходимо задать путь от корня полииерархии к текущей вершине, по которому будет произведена отрисовка. Это может быть путь, которым пользователь пришел к данной странице при навигации по структуре тезауруса, или некий «путь по умолчанию» (например, путь по остовному дереву полииерархии).

Известные алгоритмы отрисовки дерева одним запросом к реляционной БД здесь неприменимы. Однако максимальное количество запросов к БД будет равно максимальной длине пути по полииерархии тезауруса, которая, как правило, сопоставима с логарифмом от общего количества понятий, что вполне приемлемо для информационной системы. Примером реализации такого подхода может служить [18] (см. интерфейс добавления сайта в каталог).

Еще один вариант отображения полииерархии – построение остового дерева, и отображение его вышеописанными способами. В этом случае для каждого элемента тезауруса необходимо выделить из всех его предков одного, связь с которым и станет связью остового дерева (см., например, [17]). В некоторых тезаурусах заложено решение проблемы отображения именно таким путем. Специальная связь *Generic Structure* не имеет значимого семантического смысла, и служит лишь для отображения тезауруса как древовидной структуры в интерфейсах и в печатном виде [2].

Возможно также построение и отрисовка полного дерева путей по полииерархии тезауруса. Однако размер такого дерева может оказаться недопустимо большим. Например, в случае полииерархии типа «сетка рабица» (каждый элемент, кроме крайних, имеет ровно по 2 предка и по 2 потомка).

Еще один вариант отображения положения элемента в полииерархии, который будет, вероятно, полезен для пользователя – отрисовка всех соседей всех его непосредственных предков. Это будет, по сути, двухмерная таблица, легко отображаемая на экране.

Администрирование тезауруса

Интерфейсы администрирования тезауруса должны обеспечивать выполнение следующих операций:

- *Добавить новое понятие к тезаурусу.* При добавлении добавляется так же связь с некоторым другим уже существующем в тезаурусе понятием. Указывается тип этой связи.
- *Добавить связь определенного типа между понятиями.* Должно обеспечиваться ограничение: не более одной связи между двумя понятиями. При добавлении иерархической или ассоциативной связи добавляется так же парная к ней обратная связь (A ***BT*** B ⇔ B ***NT*** A; A ***RT*** B ⇔ B ***RT*** A).
- *Изменить тип связи между понятиями.* Должно обеспечиваться ограничение: Связь RT запрещена между понятиями, одно из которых является предком другого.
- *Удалить понятие и все его связи.* При удалении понятия все его потомки, не имеющие других предков, могут либо удаляться вместе с ним, либо выделяться в отдельную иерархию.
- *Удалить связь между понятиями.* При удалении иерархической связи понятие-потомок и все его потомки, не имеющие других предков, могут либо удаляться вместе с ним, либо выделяться в отдельную иерархию.
- *Добавить/изменить наиболее подходящий термин для данного понятия на некотором языке.* Должно обеспечиваться ограничение: Для каждого понятия не более одного наиболее подходящего термина на каждом языке.
- *Добавить/изменить менее подходящий термин для данного понятия на некотором языке.* При добавлении добавляется также связь к этому термину и указывается тип этой связи.
- *Добавить связь определенного типа между термином и понятием.* Должно обеспечиваться ограничение: для каждого термина не более одной связи с одним и тем же понятием.
- *Изменить тип связи между термином и понятием.*
- *Добавить/изменить комментарий к связи между термином и понятием на некотором языке.*
- *Удалить термин и все его связи.*
- *Удалить связь между термином и понятием.* Если термин не имеет других связей, он также удаляется.
- *Изменить код (идентификатор) понятия.*
- *Изменить код (идентификатор) термина.*
- *Добавить/изменить комментарий к понятию.* Должно обеспечиваться ограничение: не более одного комментария к одному понятию на одном языке.
- *Добавить/изменить комментарий к термину.* Должно обеспечиваться ограничение: не более одного комментария к одному термину на одном языке.

Интерфейсы администрирования должны включать и использовать интерфейсы просмотра тезауруса для поиска тех понятий, терминов, комментариев и связей, которые должны быть изменены. Интерфейсы редактирования могут быть также частично интегрированы в интерфейсы просмотра (в виде добавленных органов управления в окнах просмотра).

Подход к описанию тезауруса в ИСИР

Формулировка задачи

Для информационной системы ИСИР реализация тезауруса должна удовлетворять следующим свойствам:

1. Позволять хранить любые существующие тезаурусы, в частности, любые классификаторы, имеющие структуру тезауруса в соответствии со стандартами ISO 2788-1986 и 5964-1985. В том

числе, реализация должна позволять работать с многоязычными тезаурусами.
2. Позволять, используя тезаурус, индексировать ресурсы терминами данного тезауруса, а также классифицировать ресурсы понятиями тезаурусов-классификаторов. При этом работа с обоими видами тезаурусов должна осуществляться единообразно.
3. Позволять осуществлять просмотр (навигацию) по тезаурусу, а также поиск ресурсов, проиндексированных или классифицированных тезаурусом. То есть реализация должна обеспечивать эффективное выполнение необходимых для этого запросов, а именно:
   - Получить значение атрибута понятия.
   - Получить все понятия, связанные с данным понятием, связями заданных видов (для связей в соответствии со стандартами ISO или их детализаций).
   - Получить самые верхние понятия в иерархии понятий, в которую входит данное понятие.
   - Получить все термины, связанные с данным понятием, связями заданных видов (для связей в соответствии со стандартами ISO или их детализаций).
   - Получить все термины на данном языке, связанные с данным понятием связями заданных видов (для связей в соответствии со стандартами ISO или их детализаций).
   - Получить все термины на данном языке, связанные связями заданных видов (для связей в соответствии со стандартами ISO или их детализаций) с данным понятием, или с понятиями, связанными с данным понятием данными связями эквивалентности терминов на разных языках.
   - Получить значение атрибута термина.
   - Получить все понятия, связанные с данным термином, связью данного вида (для связей в соответствии со стандартами ISO или их детализаций).
   - Получить все термины, содержащие данное слово (или ключевое слово).
   - Получить полную иерархию понятий тезауруса.
   - Получить полный список терминов тезауруса.
4. Быть расширяемой. То есть допускать детализацию при необходимости некоторых связей, а так же добавление новых связей.

Описание схемы данных

Данная схема данных написана исходя из соображений, сформулированных в разделе «Тезаурусы в описании информации», а также в соответствии с вышеперечисленными требованиями. Схема данных основана на платформе RDF.

Классы модели

- *Concept.* Понятие. Имеет следующие атрибуты:
  - *ID.* Уникальный идентификатор, или код рубрики классификатора. Необязательный атрибут. Этот атрибут присутствует только тогда, когда он имеет смысловую нагрузку в тезаурусе, и не заменяет внутренний системный или технический ID записи в базе данных.
  - *IsTop.* Имеет истинное значение, если данное понятие является самым верхним в иерархии.
- *Term.* Термин. Имеет следующие атрибуты:
  - *Value.* Написание (наименование) термина на данном языке.
  - *Lang.* Язык термина
  - *ID.* Код термина. Необязательный атрибут. Этот атрибут присутствует только тогда, когда он имеет смысловую нагрузку в тезаурусе, и не заменяет внутренний системный или технический ID записи в базе данных.
- *ScopeNote.* Комментарий к понятию, термину, или отношению.
  - *Value.* Значение комментария на данном языке.
  - *Lang.* Язык комментария.
- *Concept Relation.* Реализует отношение (связь) между понятиями, снабженное комментарием.
- *Term Relation.* Реализует отношение (связь) между понятием и термином, снабженное комментарием.

Связи между объектами классов тезауруса

- Связи между понятиями. Здесь все связи являются не связями в смысле RDFS, а экземплярами класса *Concept Relation*. Каждый такой экземпляр отношения имеет связи с соответствующим объектом и субъектом данной связи *Concept Relation*.
  - *Broader Concept*
  - *Narrower Concept*
  - *Related Concept*
  - *Top Concept*
  - *Equivalent.* Для эквивалентности терминов на разных языках. Имеет подсвойства (специализации):
    - *Exact Equivalent*
    - *Inexact Equivalent*
    - *Partial Equivalent*
    - *Single To Multiple Equivalent*
    - *Non Equivalent*
- Связи между понятием и термином:
  - *Preferred Term*
  - *Nonpreferred Term.* Имеет подсвойства:
    - *Synonym.* Синоним
    - *Antonym.* Антоним
- Связи между понятием, термином или отношением, и комментарием
  - *Has Scope Note*

Данная модель имеет важное отличие от описанных ранее стандартов и моделей: Один термин может иметь связь с несколькими понятиями. Например, для одного понятия быть наиболее предпочтительным термином, а для другого – антонимом. Ограни-

чение, привязывающее каждый термин строго к одному понятию существенно в моделях, где понятия не являются отдельными объектами. В этом случае термин мог дублироваться в иерархии несколько раз, чтобы разным понятиям соответствовали заведомо разные вершины в графе представления тезауруса. Однако в модели с понятиями, как отдельными объектами, такое ограничение уже не оправдано.

Ряд тезаурусов, например, математический классификатор MSC, имеют отношения между понятиями, которые нельзя отнести строго к какому-либо из определенных в стандарте типов, либо такое отношение требует уточнения (пример см. ниже). Как правило, таких отношений в тезаурусе очень мало, а потому нецелесообразно для них выделять отдельные типы отношений. Средством описания таких отношений может стать приписывание такого отношения к одному из базовых существующих типов, с добавлением к нему комментария, характеризующего его особенности. В данной модели тезауруса любое отношение между понятиями, или между понятием и термином может быть снабжено комментарием на любом языке.

Данный подход позволит также минимизировать неминуемое дальнейшее расширение и детализацию наборов связей между терминами или понятиями, которая сейчас наблюдается в различных моделях и национальных стандартах (например, ANSI), поскольку, как альтернативу детализации, можно использовать комментарии к связям специального вида.

Ограничения модели

Ограничения этой модели включают прежде всего известные ограничения стандарта ISO. В данной модели они приобретают следующий вид (в формулах ниже ***A*** и ***B*** – понятия, ***T*** – термин):

- ***A*** *Broader Concept* ***B*** ⇔ ***B*** *Narrower Concept* ***A***.
- ***A*** *Related Concept* ***B*** ⇔ ***B*** *Related Concept* ***A***.
- ∀ ***A*** ∃! ***T***: ***A*** *Preferred Term* ***T*** (единственность для каждого языка).
- ***A*** *Related Concept* ***B*** ⇨ ***B*** не является ни предком, ни потомком ***A***.

В дополнение к этим, модель имеет еще следующие ограничения, вытекающие из предыдущих рассуждений:

- ∀ ***A*** ∃ ***B***: ***A*** *Top Concept* ***B***.
- ∃ ***A***: ***B*** *Broader Consept* ***A*** ⇨ *IsTop*(***B***) = false.

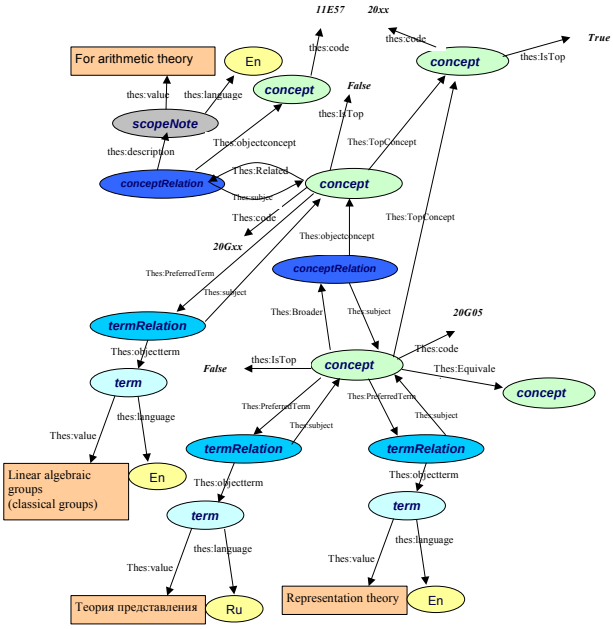
RDFS-схема

RDFS-схема для данной модели тезауруса не может быть здесь приведена по причине ограниченности объема статьи. Ее можно найти в Internet по URL http://mathnet.ru/project/rdfs/thesaurus.rdfs.

Ниже приведен пример описания одной из рубрик классификатора MSC на официальном сайте MSC http://www.ams.org/msc , а так же графически отражено представление этой рубрики и ее связей на двух языках в данной модели.

В целях экономии места из всех более узких понятий данной рубрики в реальном MSC здесь оставлено только три.

```
Term 20Gxx Linear algebraic groups (clas-
sical groups)
       RT For arithmetic theory, see
11E57, 11H56
       NT 20G05 Representation theory
```



Литература

[1] Thesaurus Construction http://instruct.uwo.ca/gplis/677/thesaur/main00.htm

[2] Thesaurus Format: Nusearch Standard Specification http://www.excavio.com/pdf/wp_nusearch_thesaurus_spec.pdf

[3] LIMBER (Language Independent Metadata Browsing of European Resources) project: http://www.limber.rl.ac.uk/

[4] A Thesaurus Interchange Format in RDF http://www.limber.rl.ac.uk/External/SW_conf_thes_paper.htm

[5] Hall, M. (2001) *CALL Thesaurus Ontology in DAML*. http://orlando.drc.com/daml/ontology/Thesaurus/CALL/

[6] RDF Thesaurus Specification http://ilrt.org/discovery/2001/01/rdf-thes/

[7] Web Thesaurus Compendium Http://www.darmstadt.gmd.de/~lutes/thesoecd.html

[8] ISO2788: Guidelines for establishment and development of monolingual thesauri, 2 nd ed., Geneva: ISO1986.

[9] ISO5964: Guidelines for establishment and development of multilingual thesauri, 1 st ed., Geneva: ISO1985.

[10] Steve Pepper, The TAO of Topic Maps http://www.ontopia.net/topicmaps/materials/tao.html

[11] Thesaurii, Techquila http://www.techquila.com/tmsinia3.html

[12] Semantic Web project
http://www.w3.org/2001/sw/
[13] Mathematical Subject Classification (MSC)
http://www.ams.org/msc
[14] Physics and Astronomy Classification Scheme (PACS) http://www.aip.org/pacs/
[15] Dewey Decimal Classification (DDC)
http://www.oclc.org/dewey/
[16] DARPA Agent Markup Language (DAML)
http://www.daml.org/
[17] Информационная система ИСИР
http://uis.isir.ras.ru
[18] Каталог ресурсов «Кирилл и Мефодий»
http://search.km.ru/url/index.asp
[19] Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М. Архитектура RDFS-системы. Практика использования открытых стандартов и технологий в системе ИСИР. 5 Всероссийская научная конференция RCDL-2003.

Approaches to description and using thesauri in information systems

A. S. Ajiev, H .M. Nguyen

The paper contains consideration of different approaches to thesauri formalization. Consideration is given also to ISO and ANSI thesauri standards. Analysis was given for some possible platforms for the formalization and description was given for peculiarities and problems involved. Also requirements are given to thesaurus realization in Semantic Web [12].

Consideration was given for peculiarities and distinctions between resource classifiers and common terminological and linguistic thesauri. Comparative analysis was given for existing RDF-based data schemes and approaches for thesauri for information systems. Consideration was given also for user and administrative interfaces and using them in information search and access.

In the second part of the paper basing on analysis done requirements are stated to thesauri description in ISIR [17]. ISIR RDFS data scheme for ISIR is stated and a small example of MSC realization in the scheme is given.