OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ТЕРМИНОЛОГИЧЕСКИЕ СЕТИ

Мальковский М.Г., Соловьев С.Ю.

Факультет ВМК МГУ имени М.В.Ломоносова, г.Москва, Россия

malk@cs.msu.su

soloviev@glossary.ru

В работе анализируется понятие терминологической сети, обсуждаются их свойства и особенности, приводится типовая технология построения таких сетей. В качестве конкретной терминологической сети описывается универсальное терминологическое пространство, использующее ограниченное количество типов бинарных отношений. Обсуждаются вопросы практического применения терминологических сетей на примере веб-ресурса www.glossary.ru; рассматриваются пользовательские сервисы, основанные на использовании свойств терминологических сетей.

Ключевые слова: глоссарий, определение, семантическая сеть, термин.

ВВЕДЕНИЕ

Человек в своих рассуждениях неизбежно оперирует терминами. Свободное владение специальной терминологией – верный признак профессионализма. Терминологические системы проблемных областей отражают соответствующий понятийный аппарат и известные отношения между понятиями. Подавляющая часть определений терминов зафиксирована в толковых словарях, глоссариях, ГОСТах и энциклопедиях. Существует и развивается терминоведение [Лейчик, 2009] – отрасль филологии, изучающая закономерности существования терминологии. В настоящей работе рассматриваются прикладные вопросы систематизации терминов и их определений в интересах развития интеллектуальных информационных технологий.

1. Структурирование терминологии

Возьмем с полки любой толковый словарь, скажем, «Новейший справочник школьника по математике» [Якушева, 2007], и откроем его на 363 странице. Среди прочих терминов на «У» здесь приводится определение угла.

> Угол – геометрическая фигура, состоящая из двух лучей с общим началом. (Здесь и далее в примерах сохранен стиль источника.)

Из приведенного определения образованный человеком легко извлекает три задействованные в нем термина: «Геометрическая фигура», «Луч» и «Начало луча». Определения первых двух терминов обнаруживаются в том же словаре на страницах 48 и 169:

> Геометрическая фигура – конечное или бесконечное множество точек.

> Луч – часть прямой, состоящая из всех точек этой прямой, лежащих по одну сторону от данной точки.

И хотя отдельного определения для «Начала луча» в словаре не обнаружено, его можно достаточно просто восстановить из контекста упоминания термина «Луч».

> Начало луча – точка на луче, относительно которой все остальные его точки расположены по одну сторону.

Результаты интеллектуальных усилий по анализу определения «Угол» можно представить в материальном виде: четыре картонные карточки с определениями, соединенные тремя наклейками-стикерами с именами бинарных отношений (см. рисунок 1).

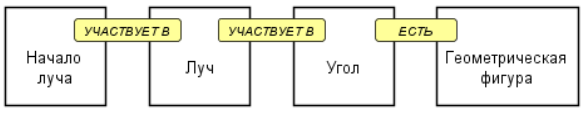

Рисунок 1 – Связи между определениями

В рассмотренном примере карточки-определения выстроились в линейную последовательность, однако так бывает не всегда. Существенно, что три определения «Луч», «Угол» и «Геометрическая фигура» имеют некоторое внутреннее устройство:

- «Луч» определяется с использованием «Начала луча»;
- «Угол» определяется с использованием «Луча»;
- «Геометрическая фигура» имеет подвид «Угол».

Наличие у определения внутреннего устройства свидетельствует о том, что соответствующий термин описывает некоторое понятие проблемной области. Фактически определение, обладающее внутренним устройством, задает типичного представителя [Нильсон, 1985] некоторого класса объектов. Наименование выявленного таким образом понятия получается переформулированием имени его типичного представителя. Как правило, переформулирование сводится к образованию множественного числа. Так, в примере из рисунка 1

- определению «Луч» соответствует понятие «Лучи»;
- определению «Угол» соответствует понятие «Углы»;
- определению «Геометрическая фигура» соответствует понятие «Геометрические фигуры».

Отметим, что в самых разных науках словосочетание «раскрыть понятие» подразумевает выявление подвидов этого понятия и описание отношений между ними с привлечением некоторого количества других концептов.

Понятия-классы и остальные определения можно рассматривать в качестве вершин семантической сети. Договоримся обозначать на рисунках понятия – овалами, а остальные определения – прямоугольниками. Набор бинарных отношений, связывающих определения, обязан содержать отношение *это-есть* (*есть*) для представления родо-видовых отношений; представительство остальных отношений определяется целью структурирования терминологии. С учетом принятых соглашений результат анализа определения «Луч» представляет собой фрагмент семантической сети, изображенный на рисунке 2.

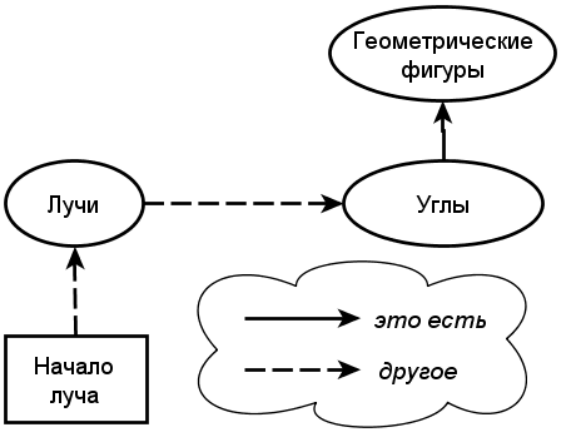

Рисунок 2 – Фрагмент семантической сети

Аналогичный анализ определения «Ломаная» порождает иной фрагмент семантической сети (рисунок 3), содержащий, в том числе, некоторое количество ранее встретившихся вершин-определений.

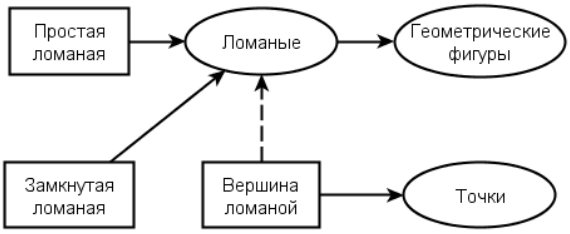

Рисунок 3 – Второй фрагмент семантической сети

Присоединение вновь построенного фрагмента (рисунок 3) к имеющейся сети (рисунок 2) порождает более сложную семантическую сеть, приведенную на рисунке 4. Идея операции присоединения состоит в том, что каждому определению в сети должна соответствовать ровно одна вершина.

Продолжая процесс построения новых фрагментов и присоединяя их к уже имеющейся сети, можно построить весьма сложную семантическую сеть, фиксирующую бинарные связи между понятиями и терминами проблемной области. Сети такого рода будем называть терминологическими сетями. Вершинам терминологической сети обязательно сопоставляются определения терминов; нарушаться это правило может только в исключительных случаях принципиального отсутствия определений. Для неопределяемых понятий в терминологической сети создается определение-заглушка.

> Терминологическая сеть – семантическая сеть, вершинами которой являются определения терминов, связанные бинарными отношениями заданных типов.

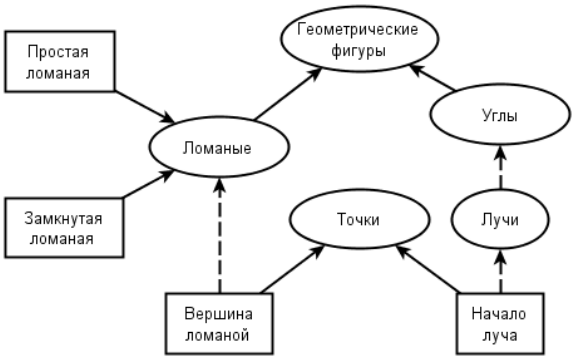

Рисунок 4 – Результат присоединения

Рассмотренная на конкретном примере технология систематизации определений, порождающая терминологическую сеть, относится к сфере интеллектуальной деятельности человека, который в этом контексте именуется научным редактором. Именно редактор в меру своей

квалификации и образования выявляет фрагменты в заданном массиве определений, а затем присоединяет их к ранее построенной сети. Вместе с тем, производительность труда научного реактора вполне поддается оценке и прогнозированию, что позволяет (хотя бы гипотетически) ставить вопрос об организации "терминологического бизнеса".

Важно, что реальная деятельность по систематизации терминологии возможна только при наличии у научного редактора адекватного программного инструментария.

2. Универсальное терминологическое пространство

В 2000 году начались работы по реализации конкретного варианта терминологической сети. Проект "Универсальное терминологическое пространство" (УТП) [Мальковский и др., 2002] проект ориентирован на интеграцию в единой сети всех терминов научной и деловой лексики. Цель проекта состоит в организации самодостаточного процесса систематизации терминологии.

По состоянию на начало 2012 года УТП содержит 53'477 определений, 9'044 из которых являются понятиями. Суммарно в УТП представлены 95'252 термина и их синонима из астрономии, биологии, географии, информатики, математики, физики, физиологии, экологии, экономической теории, внешней торговли, страхового дела, управления и из десятков других отраслей науки и видов деятельности. Между определениями УТП установлены 75'569 бинарных связей двух типов:

- отношение *это-есть* (34'566 экземпляров.); и
- отношение *относится-к* (41'003 экземпляра),

включающее в себя все типы отношений, отличающиеся от отношения *это-есть*.

Наличие только двух типов отношений объясняется необходимостью иметь технологичный программный инструмент, рассчитанный на реальные возможности научных редакторов. Отчасти это ограничение компенсируется специальными шаблонами в определениях терминов. Так, определение

> Почвенный раствор – жидкая часть почвы; вода с растворенными газами, минеральными и органическими веществами.

в явном виде информирует пользователя, что связь *относится-к*, установленная между «Почвенный раствором» и «Почвой», на самом деле является связью типа *часть-целое*.

Каждый термин, представленный в УТП, может иметь несколько синонимов и аббревиатур. определение может содержать этимологическую справка, ссылку на некоторый веб-ресурс (URL) и указатель на файл-изображение. Определение термина задается в виде линейного текста, возможно разделенного на абзацы, и/или содержащего списки перечислений.

По построению в УТП все понятия имеют различные наименования, остальные термины могут иметь «двойников». Опыт существования УТП показывает, что такое ограничение не является существенным. Как оказалось, для каждого понятия без труда находится уникальное наименование, отличающее его от других понятий собственной и других проблемных областей.

Топология текущей версии УТП с указанием наименований понятий и терминов, приписанных вершинам, находится на странице

http://www.glossary.ru/_netwrk_.htm

и доступна в формализованном электронном виде всем исследователям.

Характерная для УТП ориентация на неограниченный круг терминов отделяет от общего класса терминологических сетей собственный подкласс универсальных терминологических сетей.

> Универсальная терминологическая сеть – терминологическая сеть, предназначенная для систематизации неограниченного количества терминов.

Выдающимся, но не единственным примером универсальной терминологической сети являются клинические коды Дж. Рида [Емелин, 2000] RCC (Read Clinical Codes; Clinical Terms Version 3; CTV3). Разработка каждой универсальной сети начинается с постулирования некоторых ее свойств, которые, с одной стороны, направляют дальнейшее развитие сети, а с другой стороны – играют роль ограничителей.

3. Особенности универсальных терминологических сетей

Накопленный опыт ведения УТП позволяет говорить о некоторых принципиальных свойствах и проблемах универсальных терминологических сетей.

Проблемы редактирования | При работе с конкретной терминологией возникает большое количество ситуаций, в которых редактор вынужден принимать решения. В частности, серьезные затруднения вызывают:

- качество исходных определений;
- отсутствие определений;
- наличие альтернативных определений.

О качестве определений | Значительная часть исходных определений не выдерживает критики с точки зрения стиля и полноты; часто определения излагаются в виде адаптированном под интересы некоторой проблемной области.

Об альтернативных определениях | В некоторых науках мирно сосуществуют несколько несводимых определений одного и того же понятия: «Почва по В.В.Докучаеву», «Почва по П.А.Костычеву», «Почва по В.А.Вернадскому» и др. Каждый такой случай требует от научного редактора

индивидуального подхода для корректного представления альтернативных определений в терминологической сети.

О размере определений | Наилучшее по объему определение содержит 2-3 предложения. Наличие в определении пяти и более предложений сигнализирует о необходимости разделить текст определения между несколькими терминами.

О достоверности связей | При построении терминологической сети части терминологии, попадающей на стык проблемных областей, приходится структурировать многократно. Второе и последующие структурирования в значительной степени перепроверяют ранее проделанную работу. По результатам этих проверок можно с уверенностью утверждать, что количество ошибок в УТП мало.

О достаточности связей | В весьма редких случаях возникает необходимость представить в УТП бинарное симметричное отношений между терминами. Одно из решений этой задачи состоит в том, чтобы разместить оба термина в окрестности их родительского понятия.

Проблема недостаточности | В любом источнике терминологии всегда обнаруживается некоторое количество терминов, которые используются в определениях других терминов, но сами в этом источнике не определяются. Это обстоятельство делает неизбежным использование в процессе построения универсальной терминологической сети различных терминологических источников.

Связность | Универсальная терминологическая сеть, построенная без привлечения общенаучной, религиозной и философской терминологии, является связной. В структуре такой сети явно выделяются отрасли науки и виды деятельности, играющие объединительную роль: математика, бухгалтерский учет, страховое дело и др.

Цикличность | В терминологической сети присутствуют циклы. Как правило, существование циклов связано с различиями в ролях, которые играют крупные, сложно устроенные понятия. Типичным примером такого понятия является «Государство», которое в разных контекстах выступает и продавцом, и покупателем и регулятором рынка.

Проблема отбора | Часть терминологии, находящейся в общечеловеческом обороте, на самом деле связана с недостаточно обоснованными гипотезами. Наличие такой терминологии в универсальной терминологической сети способно серьезно исказить картину мира.

Проблема старения | В терминоведении факт старения и выхода терминологии из оборота установлен и описан. Разработка адекватных программных и изобразительных средств, автоматически поддерживающих процесс старения терминологии, составляет открытую проблему.

Проблема изменчивости | С течением времени отдельные термины («Ноосфера» и др.) претерпевают существенные изменения в своем содержании. Другими словами в разные периоды времени одному и том уже термину отвечают различные понятия. При этом изменяются и связи термина. Выявление и учет этого обстоятельства составляют отдельную, открытую проблему терминологических сетей.

Проблема мироустройства | Конструирование терминологической сети, состоящее из операций порождения и присоединения новых фрагментов аналогично восходящей схеме проектирования программ [Йодан, 1979]. Проблема лишь в конфигурации «верхних этажей» сети. Можно, скажем, полагать первичность материи или исходить из существования Бога. Как это ни странно, но вопросы мироустройства необходимо зафиксировать в качестве инженерного решения.

Инварианты | Существуют несколько инфометрических характеристик УТП, неизменных на всех этапах его существования. Эти характеристики не зависят от количества терминов в УТП и нуждаются в объяснении.

4. Веб-ресурс www.glossary.ru

Практическую востребованность концепции терминологических словарей подтверждает веб-ресурс Служба тематических толковых словарей – www.glossary.ru [Мальковский и др., 2002]. Информационное наполнение ресурса получается формальным преобразованием очередной версии УТП в структуры, удобные для построения веб-сайта. Веб-ресурс www.glossary.ru предоставляет пользователям определения терминов научной и деловой лексики, с группированные в стандартные, относительно компактные проблемно-ориентированные глоссарии.

Стандартный глоссарий определяется уникальным наименованием понятия и состоит из терминов, непосредственно связанных бинарными отношениями с соответствующей понятийной вершиной. По сути дела стандартный глоссарий содержит минимум терминологии, раскрывающий то или иное понятие.

При построении веб-ресурса существенно используется структура УТП для навигации в пространстве терминов, а также для разработки сервисов поддержки интеллектуальной деятельности. Интеллектуальные сервисы ориентированы на предоставление пользователю дополнительных возможностей по выявлению связей между терминами.

Основными интеллектуальными сервисами в проекте www.glossary.ru являются:

- сервис пополнения стандартных глоссариев;

- сервис тематической группировки;
- сервис визуализации;
- сервис формирования поисковых запросов;
- сервис веб-карт;
- сервис формирования словарей;
- сервис сортировки терминов.

Сервис пополнения позволяет включать в состав проблемно-ориентированного глоссария некоторые дополнительные термины. Например, стандартный глоссарий «Углы» не содержит определение «Начало луча», хотя этот термин и участвует в определении термина «Угол». За счет сервиса пополнения заинтересованный пользователь может получить более широкое множество терминов и их определений, включающее, в частности, и термин «Начало луча». Алгоритм пополнения [Мальковский и др., 2002] существенно использует родо-видовые связи и реализует механизм наследования свойств.

Сервис тематической группировки понятий конструирует последовательность окрестностей заданного понятия. Каждая окрестность представляет собой набор понятий, отстоящих от заданного, на известное расстояние. При группировке используются связи УТП без учета их типов. Сервис тематической группировки вскрывает «близкородственные», порой достаточно неожиданные связи между проблемными областями.

Сервис визуализации позволяет представить в наглядном виде фрагмент семантической сети, охватывающий терминологию отдельного стандартного глоссария и связанных с ним понятий [Соловьев, 2008a]. При построении графического образа существенно используется наличие в УТП ровно двух типов отношений.

Сервис интерактивного формирования поисковых запросов предоставляет пользователю многовариантную форму генерации обращений к ИПС в связи с некоторым конкретным термином. При построении формы используется содержащийся в УТП набор связей выбранного термина [Мальковский и др., 2004]. Отмеченные пользователем варианты транслируются сервисом в запрос на расширенном языке, а затем запрос передается системе Яndex на исполнение.

Сервис веб-карт представляет пользователю веб-ресурса систему терминологии в обозримом укрупненном виде [Соловьев, 2008b], и тем самым обеспечивает быструю навигацию в терминологической сети на уровне отраслей и родов. Алгоритм построения веб-карт решает задачу наглядного представления сильно разреженного графа («почти дерева»).

Сервис формирования ассоциативных терминологических словарей позволяет представить в линейном виде достаточно большую совокупность родственных терминов. Алгоритм формирования преобразует семантические связи УТП в разделы и порядок расположения терминов, а также в систему перекрестных ссылок.

Сервис сортировки терминов призван повысить «читабельность» глоссария. Существование сервиса объясняется феноменом обратной зависимости между объемом толкового словаря и привлекательностью для пользователя последовательного чтения определений. Установлено, что узкие специализированные глоссарии человек может читать почти как связный текст. Однако с ростом количества терминов и расширением проблемной области глоссарий теряет это свойство и превращается в хранилище разрозненных статей. Для иллюстрации этого феномена приведем отрывок из фундаментальной кукольной энциклопедии [Голдовский, 2004].

> Ножная кукла – перчаточная кукла, которой актер управляет с помощью ног.
> Ножной брусок – элемент управления марионетки, куда крепятся ноги куклы.
> Нос – комический кукольный народный герой традиционного голландского театра кукол.
> Носек – чешский художник-кукольник начала XX века.

Понятно, что в приведенном примере каждое следующее определение радикально изменяет контекст существования предыдущего определения, разрывая какие бы то ни было связи между соседними терминами.

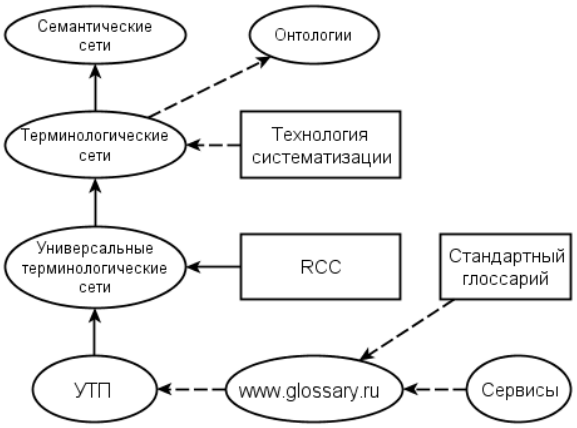

Рисунок 5 – Термин+ологическая сеть как способ представления знаний

Алфавитный принцип следования терминов противодействует образовательной функции толкового словаря. В обширных глоссариях [возникающих, например в результате пополнения] существует настоятельная необходимость в таком изменении порядка следования терминов, при котором они выстраиваются в взаимосвязанные цепочки. Вместе с тем возможный в УТП формальный переход к алфавитно-гнездовой системе расположения терминов оправдан далеко не всегда, поскольку имеющиеся семантические связи между терминами обеспечивают более высокий уровень группировки терминов. В некоторых случаях более удобными оказываются формально

построенные идеографические словари [Морковкин, 1970]. Исходя из этого, сервис сортировки терминов предлагает пользователю на выбор один из трех различных способов расположения терминов.

Приведенный перечень интеллектуальных сервисов, основанных на УТП, не является исчерпывающим. Значительный интерес для инженерии знаний представляет сервис генерации прототипов онтологий. Нерешенной остается задача применения УТП в информационном поиске вообще и в информационном поиске для проекта www.glossary.ru в частности.

Заключение

На рисунке 5 показано место терминологических сетей в системе способов представления знаний. Конечно, сети такого рода не позволяют представлять совершенно любые декларативные знания из-за ограничений на размерность отношений и из-за достаточно жесткой привязки к терминологии, однако терминологические сети существуют реально, они полезны конечным пользователям, а, кроме того, порождают и позволяют решать новые интересные задачи.

Библиографический список

[Голдовский, 2004] Голдовский, Б.П. Куклы: Энциклопедия / Б.П. Голдовский // – М.: Время, 2004. – 496 с.

[Емелин, 2000] Емелин, И.В. Интеграция стандартов медицинской информации / И.В. Емельянов // Кремлевская медицина. Клинический вестник, No. 4,- 2000. С.55-62.

[Йодан, 1979] Йодан, Э. Структурное проектирование и конструирование программ / Э. Йодан // – М.: Мир, 1979. – 415 с.

[Лейчик, 2009] Лейчик, В.М. Терминоведение: Предмет, методы, структура / В.М. Лейчик // – М.: Книжный дом «ЛИБРОКОМ», 2009. – 256 с.

[Мальковский и др., 2002] Мальковский, М.Г. Универсальное терминологическое пространство / М. Г. Мальковский, С.Ю. Соловьев // Диалог'2002. Труды международного семинара. Т 1, С. 266-270

[Мальковский и др., 2003] Мальковский, М.Г. Методы формирования глоссариев в универсальном терминологическом пространстве. / М. Г. Мальковский, С.Ю. Соловьев // Труды международной конференции “Компьютерная лингвистика и интеллектуальные технологии" – М.: Наука, 2003, С.438-440.

[Мальковский и др., 2004] Мальковский, М.Г. Структурный метод формироваия запросов к информационной системе. / М. Г. Мальковский, С.Ю. Соловьев // Труды международной конференции “Компьютерная лингвистика и интеллектуальные технологии" – М.: Наука, 2004, С.612-613.

[Мальковский и др., 2010] Мальковский, М.Г. Системы поддержки творческих процессов / М. Г. Мальковский, С.Ю. Соловьев, А.Н. Сотников // Программные продукты и инструменты, под ред. Л.Н.Королева – М.: ВМК МГУ; МАКС Пресс, 2010, No.10. С.74-85.

[Морковкин, 1970] Морковкин, В.В. Идеографические словари / В.В.Морковкин // – М.: Изд-во Моск.ун-та, 1970. – 72 с.

[Соловьев, 2008a] Соловьев, С.Ю. Образные представления терминологической сети / С.Ю. Соловьев // Сб. Прикладное программное обеспечение – М.: МИРЭА, - 2008. С.55-69.

[Соловьев., 2008b] Соловьев, С.Ю. Об одном методе генерации страниц-карт для веб-сайтов. / С.Ю. Соловьев А.Н. // Информационные процессы Том 8, No.1, 2008, С.24-39.

[Нильсон, 1985] Нильсон, Н. Принципы искусственного интеллекта / Н. Нильсон // – М.: Радио и связь, 1985. – 373 с.

[Якушева, 2007] Якушева, Г.М. Математика. Новейший справочник школьника / Г.М. Якушева // – М.: СЛОВО, Эксмо, 2007. – 479 с.

TERMINOLOGICAL NETWORKS

Malkovsky M.G., Soloviev S.Y.

Lomonosov MSU CS department, Moscow, Russia

malk@cs.msu.su

soloviev@glossary.ru

In the work the notion of terminological network and provides the technology for constructing such networks.

Introduction

In this paper we consider applied problems of systematization of terms for the development of intelligent information technologies.

Main Part

Technology systematization of terminology describes a specific example of geometric terms. The technology is based process of generating fragments of the network and the process of joining a fragment to its existing network. Semantic network with binary relations between the definitions of the terms is a result of the systematization. Such networks are called terminological networks. Terminological networks are a subclass of semantic networks.

Building a terminology network is a kind of intellectual activity. n the one hand, the work of the editor to build a terminological network is difficult. On the other hand the productivity of the editor be accurately measured. These circumstances make it possible to raise the issue of constructing a terminological network as a real business.

Software tools play a key role in building a terminological networks. Availability of software tools and standards for the construction of a terminological network impose constraints on the class of terminological networks. The "universal terminological space" is a successful example of building a large network terminology. Practical application of this project is a web-resource www.glossary.ru.

Terminological network allows us to construct non-trivial embedded application for users. Examples of intelligent services are services:

the replenishment service standard glossary,
the thematic grouping service,
the service rendering service form search queries,
the service web-card,
the service forming vocabularies,
the service sort of terms.

Conclusion

Terminological networks do not allow to represent any declarative knowledge of the restriction on the dimension of the relationship and because of the rather restricted to specific terminology, but terminology networks exist in reality and allow us to solve interesting problems.