Модел. и анализ информ. систем. Т. **17**, № **1** (2010) 5–24

УДК 025.4.06

Тезаурус как инструмент поэтологии

Бойков В.Н. [1], Захаров В.Е., Пильщиков И.А., Сысоев Т.М.

Институт космических исследований РАН,
Физический институт им. П.Н.Лебедева РАН,
Институт мировой культуры МГУ им. М.В. Ломоносова,
Межведомственный суперкомпьютерный центр РАН

e-mail: boykov_bh@bk.ru

получена 2 февраля 2010

Ключевые слова: тезаурус, рубрикатор, терминология, отношения, связи, поэтология, стиховедение

Рассмотрены принципы создания тезауруса по поэтологии — группе дисциплин, ориентированных на всестороннее теоретическое и историческое изучение поэзии. Этот тезаурус мыслится как предметно-ориентированный справочник, информационно-поисковый инструмент и инструмент аналитических исследований. Предложенные концептуальные и технологические решения оцениваются с точки зрения современных стандартов представления тезаурусов в информационных системах.

За полтора столетия, прошедших со времени издания одного из самых востребованных и не потерявших актуальности до настоящего времени тезауруса Роже [1], в мире появилось необозримое число тезаурусов, как общеязыковых, так и охватывающих узкие, специальные области знания. В последнее время интерес к тезаурусам существенно возрос в связи с развитием систем семантической обработки информации и в связи с ключевой их ролью в отражении онтологии конкретных областей знания. По существу, тезаурус представляет собой словарь концептов с иерархической структурой и набором парадигматических отношений, указывающих на общность или противопоставление значений лексических (терминологических) единиц и на особенности использования дескрипторов.

Основные недостатки, отмечаемые в разработанных тезаурусах, носят преимущественно субъективный характер: это и неполнота соответствующей концептуальной лексики, и узкий спектр семантических (ассоциативных) связей, и произвол в выборе дескрипторов в синонимических гнездах, часто при отсутствии иноязычных эквивалентов [2, с. 44]. Как следствие, при информационном поиске при помощи разных тезаурусов, родственных по предметной области, результаты в значительной

[1]Работа поддержана Российским фондом фундаментальных исследований (проект № 07-06-00156)

степени могут не совпадать. Вместе с тем указанные недочеты говорят о наличии объективных предпосылок для их устранения.

Авторы статьи ставят своей целью создание предметно-ориентированного тезауруса поэтологии (ТП). Под поэтологией мы будем понимать группу дисциплин, ориентированных на всестороннее теоретическое и историческое изучение поэзии. Сюда входят следующие дисциплины: 1) теория и история стиха (анализ и описание метрики, ритмики, строфики, рифмы, морфологии и синтаксиса стихотворной речи); 2) теория и история поэтического языка (анализ и описание поэтической фоники и просодии, лексики и фразеологии, словообразования и др. уровней поэтического языка); 3) поэтика, риторика и стилистика стихотворного текста; 4) сюжетология, мотивный анализ, нарратология (исследование приемов художественного повествования), теория и история поэтических жанров; 5) история национальной поэзии (изучение генезиса и эволюции поэтических форм; хронология и периодизация литературного процесса; история художественных направлений, школ и групп; изучение биографии поэтов); 6) библиография национальной поэзии.

Рассматриваемый в данном случае ТП в качестве тезауруса верхнего уровня предполагает тезаурус по филологии в целом, из которого должна привлекаться по необходимости лексика других филологических наук. Таким образом, процесс создания тезаурусов как верхнего, так и нижнего уровня должен носить итерационный характер. Одной из особенностей ТП является то, что в области его применения находятся как поэтические произведения, так и поэтологические исследования. Являясь частью лексики последних, ТП по отношению к поэтическим текстам может рассматриваться как метаязык.

Предназначение тезауруса как классификатора определенной проблемной области имеет, в основном, два аспекта: во-первых, это предметно-специфицированный справочник, во-вторых, основа системы информационного поиска. Вместе с тем тезаурус может рассматриваться в качестве инструмента аналитических исследований. ТП — это терминологический справочник, информационно-поисковый инструмент, основа для аналитической спецификации поэтического произведения, его стихового и языкового строения.

В соответствии с тематическим профилем различают межотраслевые, отраслевые и узкотематические тезаурусы. ТП следует отнести к узкотематическим с ожидаемым объемом терминов не более 2 тыс. единиц. В тезаурус попадает терминология 4 типов (Т1–Т4). Терминология тематики тезауруса (Т1) должна отражать лингвостиховедческие понятия (т. е. понятия, используемые при описании метрики, ритмики, строфики, рифмы и фоники, морфологии и синтаксиса стиха); понятия теории литературы, применимые к русской поэтической практике и используемые для описания видов художественной речи, поэтических жанров и проч.; понятия истории русской поэзии, используемые для описания поэтических стилей, направлений, течений, школ и т. д. В качестве терминологии смежных областей (Т2) должны быть представлены понятия литературоведения (общей теории литературы, сравнительного литературоведения, истории русской и мировой литературы), фольклористики, литературной критики, поэтики, риторики и лингвистики (фонетики, фонологии, лексикологии, морфологии и синтаксиса и некоторых других дисциплин). Кроме того, в тезаурус должны входить общенаучная лексика (Т3) и общеязыковая

лексика (Т4) [3, с. 22–24].

В современных исследованиях поэтического творчества (в главной задаче поэтологии) вполне отчетливо различаются неформальные и формальные аспекты. Формальный подход является основным при изучении специфики поэтического произведения и его стихового и языкового строения. В связи с этим возникают многочисленные задачи разметки параметров стиха, таких как метро-ритмические характеристики (стихотворный размер, вид клаузулы, характеристика рифмовки, наличие тех или иных ритмических особенностей и т. д.); параметров произведения в целом, относящихся к его виду, жанру, композиционной форме, и прочим формальным признакам. Номенклатура указанных параметров и представляет собой формальную спецификацию стиха. При этом важную роль должна играть систематизация представлений о комплексах характеристик, специфицирующих стих и поэтическое (стихотворное) произведение на основе ТП. Такая обобщенная спецификация (формализованная система специфических признаков) как стиха, так и произведения позволит определить направления исследований и комплекс аналитических задач поэтологии.

Научные издания бывают (хотя далеко не всегда) снабжены достаточно обширным справочным аппаратом, куда входят различные указатели: именные, географические, хронологические, библиографические и некоторые другие. Они являются своеобразным путеводителем по конкретному изданию и определенным образом его специфицируют. В спецификации поэтологического исследования немаловажную роль может сыграть ТП: частотный предметный указатель, составленный на его основе, может служить не только навигатором в научной работе, но может дать значительно большее представление о направлении и конкретике данного исследования, чем система других указателей, включая библиографический набор ключевых слов и оглавление.

Важным аргументом в пользу рассматриваемого подхода является достаточно хорошая теоретическая и методическая обеспеченность процесса создания тезаурусов, наличие нормативных документов, регламентирующих этот процесс, и богатый практический опыт, накопленный как у нас в стране, так и за рубежом (подробнее см. ниже). Однако до сих пор попытки создания литературоведческих тезаурусов не давали успешных результатов, например: [4] (устаревший с точки зрения состава терминов, структуры и дефиниций тезаурус, являющийся, однако, первым опытом тезаврирования литературоведческих терминов) и [5] (новый словарь-тезаурус, рассчитанный на школьников и ориентированный преимущественно на разработку словарной, а не тезаурусной части). В области же стиховедения такая задача даже не ставилось (виной тому, в частности, слабая разработанность общей теории литературы и общей теории стиха, в результате чего разные ученые дают не просто различные, а принципиально не сопоставимые определения рифмы, стопы или, скажем, строфы). Существует, однако, множество предшествующих исследований и разработок по тезаурусам, из которых можно выделить такие тематически и методологически близкие поставленным задачам, как Тезаурус по теоретической и прикладной лингвистике [6], Экспериментальный системный толковый словарь-тезаурус стилистических терминов [7], Информационно-поисковый тезаурус ИНИОН по языкознанию [8], а также исследования [9]–[13].

Проектируемый тезаурус должен охватить около 1000 терминов, используемых в следующих поэтологических дисциплинах: (1) стиховедение; (2) стилистика; (3) поэтика; (4) риторика; (5) история литературы; (6) герменевтика; (7) теоретические школы и направления; а также употребляемые в поэтологических исследованиях термины, относящиеся к (8) логике и методологии науки. На основе анализа терминологического материала и специфики его использования в различных филологических дисциплинах разработана следующая классификация (рубрикация) предметной области:

1. **Стиховедение**:
   - 1.1 Стих:
     - 1.1.1 Метрика:
       - 1.1.1.1 квантитативная метрика;
       - 1.1.1.2 силлабика;
       - 1.1.1.3 силлабо-тоника;
       - 1.1.1.4 тоника;
       - 1.1.1.5 свободный стих (верлибр);
       - 1.1.1.6 маргинальные системы стихосложения;
     - 1.1.2 Ритмика:
     - 1.1.3 Строфика:
       - 1.1.3.1 строфы;
       - 1.1.3.2 квази- и гиперстрофические формы;
       - 1.1.3.3 твердые формы;
     - 1.1.4 Рифма:
       - 1.1.4.1 типы рифмы по количеству слогов;
       - 1.1.4.2 типы рифмы по фонетическому составу;
     - 1.1.5 Лингвистика стиха:
       - 1.1.5.1 фоника стиха;
       - 1.1.5.2 морфология стиха;
       - 1.1.5.3 синтаксис стиха;
   - 1.2 Проза (в отличие от стиха):
     - 1.2.1 формы прозы;
     - 1.2.2 членение прозы;
2. **Стилистика**:
   - 2.1 функционально-стилистические разновидности языка;
   - 2.2 функции языка;
   - 2.3 уровни языка:
     - 2.3.1 фоника и просодия;

- 2.3.2 лексика и фразеология;
- 2.3.3 словообразование;
- 2.3.4 синтаксис.

3. **Поэтика**:
   - 3.1 Теоретическая поэтика:
     - 3.1.1 мотивы и сюжеты;
     - 3.1.2 композиция;
     - 3.1.3 нарратология (анализ повествования);
     - 3.1.4 роды, виды, жанры;
   - 3.2 Историческая поэтика.
4. **Риторика**:
   - 4.1 стили речи;
   - 4.2 тропы и фигуры.
5. **История литературы**:
   - 5.1 Фольклор (устное народное творчество);
   - 5.2 Письменная (авторская) литература:
     - 5.2.1 хронология и периодизация;
     - 5.2.2 направления, школы, группы;
     - 5.2.3 биография писателя.
6. **Герменевтика**:
   - 6.1 интерпретация;
   - 6.2 комментарий.
7. **Теоретические школы и направления**:
   - 7.1 культурно-историческая школа;
   - 7.2 мифологическая школа;
   - 7.3 психологическая школа;
   - 7.4 психоанализ;
   - 7.5 марксизм;
   - 7.6 формализм (формальный метод);
   - 7.7 новая критика;
   - 7.8 феноменология;
   - 7.9 структурализм (структурно-семиотические методы);

7.10 постструктурализм;

7.11 гендерные исследования;

7.12 другие направления (генетическая критика, генеративный подход, семиотика культуры, культурная антропология и т. д.).

8. **Логика и методология науки**:

8.1 процедуры;

8.2 методы исследования;

8.3 приемы исследования.

В ТП указанная рубрикация реализуется двояко: во-первых, как иерархическая структура (“дерево”) системы терминов, на промежуточных и терминальных узлах которого расположены статьи тезауруса; во-вторых, как система связей, образуемая содержанием поля “рубрика” (то есть отнесенность к рубрике) в каждой конкретной статье тезауруса. Для эффективной работы с тезаурусом необходимо обеспечить переходы от ссылок в поле “рубрика” к соответствующим узлам “дерева” системы. Таким образом, формальной моделью тезауруса является граф; узлы графа — это понятия и связанная с ними информация.

Под системой понятий, отражаемой в тезаурусе, понимается упорядоченный перечень “терминов” (терминологических слов и устойчивых терминологических словосочетаний данной предметной области) с комплексом информации, характеризующей и каждый “термин”, и отношения между ними. Было принято решение ограничиться при описании терминов учетом вхождения их в следующие пять типов отношений: (1) синонимия; (2) родо-видовые отношения; (3) отношения части и целого; (4) отношение смежности; (5) ассоциативные (свободные) отношения. Кроме того, дифференцируя омонимы, тезаурус de facto учитывает еще один тип отношений: (6) омонимия.

В рамках ТП разработана структура статьи тезауруса и определены правила ее заполнения. Каждая статья тезауруса состоит из 18 полей (обязательными для заполнения являются поля “термин”, “определение” или “синонимы”, “дисциплина” и “рубрика”). Выделены следующие поля:

1. **“Термин”.** Слово или словосочетание на русском языке или на иностранном языке (греческом, латинском, французском и т.д.) — в том случае, если в качестве единственного или основного варианта термина принят иноязычный вариант.

2. **“Варианты написания”** (для терминов с неустойчивой орфографией или несколькими традициями употребления).

3. **“Этимология”.** Происхождение слова (слово языка источника, сведения о его значении и внутренней форме, иногда о его дериватах).

4. **“Иноязычные эквиваленты”.** Сюда попадают эквиваленты тезаврируемого термина на иностранных языках, в которых существует устойчивая традиция его употребления и литература по темам, предполагающим использование данного термина.

5. **“Синонимы”.** Сюда включаются эквиваленты тезаврируемого термина в русском языке. Синонимия может быть полной и неполной (дискуссионным вопросом для ряда терминов будет включение их в 5-е или 14-е поле тезаврируемого термина). Для малоупотребительных терминов содержание поля “Синонимы” может исчерпывать дефиницию, оставляя поле “Определение” пустым.

6. **“Определение”.** Дефиниция, описание термина, текст на русском языке. Опыт показывает, что дефиницию приходится почти всегда компилировать заново: заимствовать ее из одного из источников тезауруса обычно не удается в силу разных принципов определения термина в тезаурусе и его источниках. Расхождения между предлагаемой дефиницией и иными, существующими в научной традиции, учитываются в поле 7.

7. **“Альтернативные определения”.** Для одного термина может быть указано несколько дефиниций. Существует практика вынесения их за пределы статьи тезауруса (в комментарий и т. д.). Представляется, однако, что для тезауруса как средства информационного поиска такие определения должны включаться в основной состав статьи.

8. **“Аннотации”.** Принято решение помещать в это поле гиперссылки на полные тексты статей, послужившие основными источниками статьи тезауруса. Для печатной версии статьи могут быть использованы эксцерпты из этих источников.

9. **“Родовое понятие”.** Экспликация родо-видовых отношений термина (отношение, обратное отношению, фиксируемому в поле 10). У термина может быть только одно родовое понятие. Если их оказывается несколько, мы имеем дело с омонимами, и статья должна быть разбита на две.

10. **“Видовые понятия”.** Экспликация родо-видовых отношений термина (отношение, обратное отношению, фиксируемому в поле 9). У термина может быть несколько видовых понятий (в ряде случаев список видовых понятий является открытым).

11. **“Целое”.** Целое, к которому относится “часть”, определяемая тезаврируемым термином (отношение, обратное отношению, фиксируемому в поле 12).

12. **“Компоненты”.** Компоненты, части “целого”, определяемого тезаврируемым термином (отношение, обратное отношению, фиксируемому в поле 11).

13. **“Смежность”.** Фиксация метонимических отношений термина.

14. **“Ассоциации”.** Все прочие термины, связанные с тезаврируемым термином, отношения с которыми не определяются более подробно.

15. **“Дисциплина”.** В описанной выше классификации (рубрикации) — рубрика первого уровня.

16. **“Рубрика”.** В описанной выше классификации (рубрикации) — конкретная рубрика, в которую помещена данная статья тезауруса.

17. **“Источники информации”.** В качестве источника определения и содержания других полей тезауруса выступают справочно-энциклопедические и словарные издания из предварительно отобранного списка. В основные источники информации могут попадать не все издания из этого списка.

18. **“Дополнительные источники информации”.** Все источники, использованные для составления статьи тезауруса и не перечисленные в поле 17.

Примеры заполнения полей в конкретных статьях ТП см. в *Приложении.*

ТП разрабатывается, с одной стороны, как самостоятельный продукт, а с другой — как важнейшая составляющая информационно-аналитической системы по русской поэзии ([3]; ср. [14]). К настоящему моменту выработан ряд стандартов на представление тезаурусов в информационных системах. Наиболее известными среди них являются ISO 5964:1985 — руководство по построению и разработке многоязычных тезаурусов [15], ISO 2788:1986 — руководство по построению и разработке одноязычных тезаурусов [16], ГОСТ 7.24–90 — тезаурус информационно-поисковый многоязычный [17], ГОСТ 7.25–2001 — тезаурус информационно-поисковый одноязычный [18], ANSI/NISO Z39.19–2005 — руководство по построению, структурированию и обслуживанию одноязычных управляемых словарей [19]. Перечисленные стандарты описывают тезаурус похожим образом — как набор **терминов (лексических единиц)**, связанных между собой различными **отношениями**. При этом, в зависимости от стандарта, специфицируются различные допустимые атрибуты (свойства) у терминов, допустимые связи, а также ограничения, налагаемые на термины и связи между ними.

Типичными связями между терминами, которые определяются в данных стандартах, являются:

USE, см. (смотри) — используется для того, чтобы связать менее предпочтительный термин (аскриптор) среди множества терминов, определяющих одно и то же понятие, с наиболее предпочтительным (дескриптором);

UF (used for), с (синоним) — обращение связи USE. С помощью данной связи предпочтительный термин для какого-либо понятия связывается с менее предпочтительными терминами;

BT (broader term), в (выше) — связь с более общим термином в иерархии;

NT (narrower term), н (ниже) — связь с более специализированным термином.

Типы связей между терминами сами, в свою очередь, могут образовывать иерархию. Например, связь BT в некоторых стандартах имеет несколько специализаций (уточнений), таких как BTG (broader term generic), BTI (broader term instance) и BTP (broader term partial), которые обозначают, соответственно, отношения “род-вид”, “класс-экземпляр” или “целое-часть”. При этом, если между терминами существует специализированное отношение, то считается, что между ними установлено

и более общее отношение. Например, если некоторые термины A и B связаны отношением BTI, то, как следствие, они связаны отношением BT.

Опишем связи и свойства терминов, используемых в модели-прототипе ТП, и их соответствие стандартным подходам.

Варианты написания. Несмотря на сходство с отношением USE/UF, варианты написания не являются в строгом смысле синонимами, а скорее представляют уточнение этой связи. Поэтому "вариант написания" следует выделить либо в отдельное отношение по аналогии с USE/UF, либо реализовать как его подтип.

Иноязычные эквиваленты. Данное отношение можно поддержать либо как свойство термина, либо как связь с терминами на иностранном языке. Второй подход более предпочтителен, поскольку позволяет явно указать язык иностранного эквивалента и реализовать единообразный поиск по терминам независимо от их языка. В то же время основным языком тезауруса является русский, и альтернативной иерархии иностранных терминов не строится (по крайней мере, пока), что должно быть учтено в пользовательском интерфейсе.

Синонимы имеют такое же значение, как и в стандартах.

Определение достаточно точно соответствует свойству "Scope Note" (уточнение значения и области применения), и использование этого свойства для хранения определений является оправданным.

Родовое понятие/видовые понятия соответствуют отношениям BTG/NTG.

Целое/части соответствуют отношениям BTP/NTP.

Ассоциации соответствуют отношению RT.

Этимология написания, альтернативные определения, источники информации и **дополнительные источники информации** являются свойствами термина, специфическими для ТП.

Отдельного внимания заслуживает визуальная иерархия (рубрикатор) ТП. В имеющемся прототипе он представлен в виде дерева категорий, которые задают структуру, без собственных свойств. Но, в то же время, имеющиеся рубрики можно рассматривать как термины, образующие иерархию с помощью свойств BT/NT. Для того, чтобы обеспечить визуальное отличие терминов от рубрик, целесообразно ввести новое свойство для термина, характеризующее термин как рубрику или как обычный термин. Для выделения основной визуальной иерархии в случаях, когда у одного термина есть несколько более общих в различных смыслах, предполагается использовать отношения структуры (GS). В частности, это облегчит навигацию по связям BT/NT среди терминов, которые можно будет отображать в общей иерархии, в отличие от текущего решения, когда термин автоматически является конечной вершиной.

Описанная выше модель ТП еще не сформирована окончательно, и в дальнейшем, в том числе и в процессе его наполнения, может быть принято решение о целесообразности добавления дополнительных свойств или отношений. К примеру, уже на данном этапе видна потенциальная польза от включения связи **"антоним"**, ее частными случаями являются отношения **"противоречие"** и **"противоположность"**, которые могут также потребовать различения. Рассматривается вопрос о добавлении такого уточнения иерархических связей, как **"система/компонент"**. По мере наполнения в тезаурусе будут появляться новые рубрики; возможно, в том

числе, появление новых рубрик первого уровня (например, **“библиография”**).

Рассмотрим различные подходы к описанию тезаурусов на основе **RDF**. Стандарт RDF (Resource Definition Framework) [20] определяет способ представления информации, удобный для машинной обработки, и разработанный для применения в Интернете. Особенности данного языка легко описать с помощью сравнения с широко распространенным языком XML. В то время как XML определяет представление, структуру информации, RDF описывает ее семантику (смысл) и может быть представлен в различных формах, одной из которых является XML. Представление данных в RDF виде позволяет компьютеру “понимать” информацию (в степени, определенной применяемыми алгоритмами). Одной из целей создания такого языка было обеспечить возможность обработки данных независимо от приложения, с помощью которого эта информация была получена. Все эти свойства делают язык RDF естественным выбором для представления структурированной информации в Интернете. Кроме того, RDF может быть использован и внутри системы в качестве одной из моделей данных.

Основной конструкцией языка RDF является утверждение, что некоторый объект обладает заданным свойством. Из множества таких утверждений строится вся представляемая в RDF информация. При этом в стандарте определяется, каким образом идентифицируется объект, о котором делается утверждение, и свойство, которое приписывается объекту. Значением свойства могут быть как простые типы (такие, как число или строка), так и сложные типы (в частности, ссылки на другой объект). Смысл утверждений определяется свойством, поэтому для возможности обработки данных различными приложениями есть необходимость в общих “словарях свойств”. В свою очередь, для описания таких “словарей свойств” разработаны специальные стандарты — OWL (Web Ontology Language), OWL2 [21][22].

В статьях [23][24][25] рассмотрены следующие схемы данных на основе RDF, предназначенные для описания тезаурусов в информационных системах.

LIMBER (Language Independent Metadata Browsing of European Resources) [26]. “Данный формат изначально разрабатывался для многоязычного тезауруса ELSST (European Language Social Science Thesaurus). Однако в настоящий момент LIMBER предлагает данную модель как универсальную, для представления многоязычных тезаурусов. <...> Эта модель хорошо подходит для описания тезаурусов, в которых существуют разные иерархии терминов на разных языках. Однако в ней язык термина является атрибутом понятия, а не термина. Как следствие, такая модель не пригодна для описания многоязычных классификаторов ресурсов, в которых понятия семантически не связаны с каким-либо определенным языком” [24].

ILRT (Institute for Learning & Research Technology, University of Bristol) [27]. Эта модель “строилась в расчете на работу не только с тезаурусами в обычном, "лингвистическом" смысле, но и с классификаторами. Потому язык термина в этой модели 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 <...> По своей сути эта модель предназначена для одноязычных тезаурусов и тезаурусов-классификаторов, поскольку механизм полной поддержки многоязычных тезаурусов никак не прописан, а обозначено только направление, как

это можно сделать в рамках данной модели" [24].

DRC (Dynamics Research Corporation) [28]. "Эта модель наиболее точно соответствует модели одноязычного тезауруса ISO 2788:1986. В частности, в ней отсутствует класс понятий, и все связи существуют только между терминами. Некоторые связи уточнены, в частности выделены разные виды связей менее предпочтительными терминами. Модель реализована на языке DAML. <...> Поскольку в модели нет понятий как отдельных объектов, она не удобна для реализации классификаторов" [24].

Создаваемый тезаурус, как и любой другой, должен стать развивающейся структурой, и его никогда нельзя считать вполне законченным, то есть его содержание, объем и форма должны постоянно корректироваться в соответствии с изменениями, происходящими в обслуживаемой им тематической области. В связи с этим должен быть предусмотрен специальный технологический процесс ведения тезауруса. Основным способом доступа к информационной системе ТП предполагается доступ с использованием Интернета и применением стандартных средств и протоколов (HTML, HTTP). Для большинства сценариев использования — таких, как поиск информации, просмотр или изменение информации по определенному термину, — подобная форма доступа является достаточно удобной и привычной. Однако для ряда задач он не подходит. Прежде всего среди таких задач следует отметить массовую загрузку или выгрузку информации, а также различные аналитические задачи, которые требуют выполнения запросов, допустимых в применяемой модели данных, но не предусмотренных в пользовательском интерфейсе.

Для решения этой проблемы в информационной системе предполагается возможность загрузки и выгрузки специфицированной части общего массива данных в следующих форматах:

RDF. Как уже было упомянуто выше, RDF является способом представления данных, предназначенным для машинной обработки. Поскольку он лежит в основе модели данных ТП, возможность выгрузки или загрузки данных в этом формате является вполне естественной. Информация в данном формате представлена наиболее полно и точно, но недостаточно удобна для задач редактирования и, в особенности, наполнения тезауруса.

Текстовые файлы в форматах офисных приложений. Данный формат наиболее удобен для пользователей, наполняющих тезаурус. Вместо непосредственной работы с информационной системой, которая, в частности, требует постоянного интернет-доступа, авторы имеют возможность выгрузить часть описаний терминов в текстовый файл, произвести необходимые изменения (в том числе добавить новые термины по аналогии с уже имеющимися) и загрузить измененные описания в систему, когда будет возможность.

Текст в данном формате следует определенным соглашениям для того, чтобы его можно было разобрать с помощью компьютера. Как видно из примеров, приведенных в *Приложении*, описание термина состоит из ряда параграфов, каждый из которых может начинать определение нового свойства или связи термина. Начало определения нового свойства или связи идентифицируется по порядковому номеру и названию свойства. (Таким образом, текст свойства может включать более одного параграфа.)

При загрузке данных как из формата RDF, так и из текстового формата, информационной системой тезауруса выполняется проверка на корректность входных данных. Среди прочего проверяется наличие всех обязательных свойств термина, корректность ссылок на другие термины (то есть присутствие их в базе), и то, что все указанные свойства известны (добавление неизвестного свойства, в том числе из-за ошибки в написании, не позволит выполнить загрузку).

Помимо стандартных символов, представимых в кодировке Unicode, для ТП требуется поддержка значков стиховедческой нотации. В общем виде задача ставится как возможность поддержки пополняемого списка специальных символов, которые могут быть использоваться в тексте наравне с обычными символами.

Для отображения специальных символов вместе с текстом в браузерах применяются различные способы, прежде всего использование специальных шрифтов, изображений или приложений. При этом наибольшей совместимостью обладает метод отображения с помощью встроенных в текст изображений (images). Недостатками этого метода является относительная сложность выравнивания изображений в соответствии с текстом, а также масштабирования изображения при смене размера текста. Кроме того, пользователи, у которых не отображаются картинки (например, работающие на текстовых терминалах либо отключившие показ картинок для экономии трафика), не будут видеть эти символы. Однако применение специальных шрифтов или встраиваемых приложений в большинстве случаев требует от пользователя специальной настройки.

Кроме того, в языке HTML предусмотрены средства, которые позволяют вместо картинки (в тех случаях, когда картинка не отображается) поместить замещающий текст.

Помимо отображения символов, стоит задача их набора. В текстовых документах, полученных в результате экспорта или предназначенных для импорта, поддержку специальных символов необходимо сделать, не выходя за рамки обычного текста. Хотя в базе данных и в RDF файлах имеются дополнительные возможности по сравнению с текстовыми документами, для упрощения обработки и совместимости различных форматов имеет смысл поддерживать специальные символы таким же образом, как и в текстовых документах.

Для данной задачи достаточно естественным вариантом решения является экранирование (escaping) символов. Данный метод заключается в том, что определенный символ (чаще всего символ ‘\’) объявляется специальным, после которого ожидается символ или множество символов, интерпретируемых особым образом. Для хранения в тексте собственно символа экранирования он дублируется (пример: ‘\\’). Данный метод позволяет поддержать произвольное количество поэтологических символов, присвоив им имена или номера, удобные для пользователя.

Итак, концептуализация предметной области русской поэзии предполагает прежде всего развитие и систематизацию понятийного аппарата теории литературы, истории литературы, теории и истории русского стиха. Наиболее адекватным механизмом представления знаний по русской поэзии представляется предметно-ориентированный тезаурус, а одной из насущных задач по реализации такого тезауруса — разработка специального технологического процесса его ведения.

Список литературы

1. Roget P. M. Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition. London: Longman, 1852.

2. Гиляревский Р. С., Шашкин А. В., Белозеров В. Н. Рубрикатор как инструмент информационной навигации. СПб.: Профессия, 2008.

3. Захаров В. Е., Бойков В. Н., Вигурский К. В., Пильщиков И. А. Русская поэзия: проблемы консолидации и анализа в электронном формате. М.: Пробел-2000, 2004.

4. Орлов А. Н. Тезаурус информационно-поисковый по литературе, литературоведению, фольклору и фольклористике. М.: ИНИОН АН СССР, 1975.

5. Русова Н. Ю. От аллегории до ямба : Терминологический словарь-тезаурус по литературоведению. М.: Флинта; Наука, 2004.

6. Никитина С. Е. Тезаурус по теоретической и прикладной лингвистике. М.: Наука, 1978.

7. Никитина С. Е., Васильева Н. В. Экспериментальный системный толковый словарь стилистических терминов: Принципы составления и избранные словарные статьи. М.: ИЯз РАН, 1999.

8. Смиренский В. Б. Языкознание. Информационно-поисковый тезаурус ИНИОН. М.: ИНИОН РАН, 2007. — С вкладкой на CD-ROM.

9. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах. М.: Наука, 1989.

10. Никитина С. Е., Васильева Н. В. Термины лингвистической поэтики в словаре тезаурусного типа // Славянский стих: Стиховедение, лингвистика и поэтика. М.: Наука, 1996. С. 50–57.

11. Рубашкин В. Ш., Лахути Д. Г. Семантический (концептуальный) словарь для информационных технологий // Научно-техническая информация. Сер. 2: Информационные процессы и системы. 1998, № 1. С. 19–24; 1999, № 5. С. 1–12; 2000, № 7. С. 1–9.

12. Мдивани Р. Р. О разработке серии тезаурусов по социальным и гуманитарным наукам // Научно-техническая информация. Сер. 2: Информационные процессы и системы. 2004. № 7. С. 1–9.

13. Смиренский В. Б. Ключевые слова, дескрипторы и концепты в тезаурусе по общественным наукам. — Режим доступа: www.dialog-21.ru/dialog2006/materials/html/ Smirenskiy.htm

14. Jing Y., Croft W. B. An Association Thesaurus for Information Retrieval // RIAO 94 [4th International Conference "Recherche d'Information Assistée par Ordinateur"]: Conference Proceedings. New York: Rockefeller University, 1994. P. 146–160.

15. ISO 5964:1985. Guidelines for the establishment and development of multilingual thesauri.

16. ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri.

17. ГОСТ 7.24-90. СИБИД. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению.

18. ГОСТ 7.25-2001. СИБИД. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления.

19. ANSI/NISO Z39.19 — 2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.

20. Resource Description Framework (RDF) // www.w3.org/RDF/

21. Bechhofer S., van Harmelen F., Hendler J., Horrocks I., McGuinness D. L., Patel-Schneider P. F., Stein L. A. OWL Web Ontology Language: Reference. — Режим доступа: www.w3.org/TR/owl-ref/

22. W3C OWL Working Group. OWL 2 Web Ontology Language: Document Overview — Режим доступа: www.w3.org/TR/owl2-overview/

23. Nguyen Manh Hung. Thesaurus Implementation in Integrated System of Information Resources (ISIR) // Programming and Computing Software. 2004. Vol. 30, Issue 4. P. 230–240.

24. Нгуен М. Х., Аджиев А. С. Описание и использование тезаурусов в информационных системах, подходы и реализация // Электронные библиотеки. 2004. Вып. 1. — Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part1/NA

25. Matthews B. Review of RDF Thesaurus Work: A review and discussion of RDF schemas for thesauri. — Режим доступа: http://www.w3.org/2001/sw/Europe/reports/thes/8.2/

26. Matthews B. M., Miller K., Wilson M. D. A Thesaurus Interchange Format in RDF. — Semantic Web Conference 2002, доклад.

27. Cross P., Brickley D., Koch T. RDF Thesaurus Specification (draft). — Режим доступа: www.ilrt.org/discovery/2001/01/rdf-thes/

28. Hall M. CALL Thesaurus Ontology in DAML. — Dynamics Research Corporation, 2001.

Приложение. Примеры статей ТП

1. **термин** дольник
2. **варианты написания**
3. **этимология** от рус. дольный (по аналогии с трехдольник, четырехдольник, с одной стороны, и ударник, фразовик и т. п., с другой)
4. **иноязычные эквиваленты** *англ.* dolnik
5. **синонимы** паузник, паузный трехдольник
6. **определение** Стих, в строках которого число иктов (метрических ударений) выравнено по количеству, а между ударениями располагается один-два безударных слога.
7. **альтернативные определения**
8. **аннотации (статьи)** Ахманова; Краткая литературная энциклопедия; Словарь литературоведческих терминов
9. **родовое понятие** размер стихотворный
10. **видовые понятия**
11. **целое (к которому относится термин)** тоника
12. **компоненты (части)**
13. **смежность**
14. **ассоциации** акцентный стих, гексаметр, переходный размер, тактовик
15. **дисциплина (рубрика первого уровня)** стиховедение
16. **рубрика** метрика : тоника
17. **источники информации** *Ахманова О. С.* Словарь лингвистических терминов. М.: Сов. энцикл., 1966; *Гаспаров М. Л.* Дольник // Большая советская энциклопедия. — 3-е изд. — М.: Сов. энцикл., 1972. — [Т.] 9: Дебитор — Евкалипт; *Ермилова Е. В.* Дольник // Краткая литературная энциклопедия. — М.: Сов. энцикл., 1964. — Т. 2: Гаврилюк — Зюльфигар Ширвани. — Стб. 734—735; *Карпов А.* Дольник // Словарь литературоведческих терминов / Ред.-сост. Л. И. Тимофеев и С. В. Тураев. — М.: Просвещение, 1974. — С. 69—70.
18. **дополнительные источники информации** *Зунделович Я.* Дольники // Литературная энциклопедия: Словарь литературных терминов: В 2-х т. / Под ред. Н. Бродского, А. Лаврецкого, Э. Лунина и др. — М.; Л.: Изд-во Л. Д. Френкель, 1925. — Т. 1: А—П. — Стб. 213—214; *Квятковский А. П.* Дольник // Квятковский А. П. Поэтический словарь. — М.: Сов. энцикл., 1966. — С. 107.

1. **термин** рифма
2. **варианты написания**
3. **этимология** от *греч.* ῥυθμός ‘размеренность; соразмерность’
4. **иноязычные эквиваленты** *англ.* rhyme, *англ.* rime, *франц.* rime, *исп.* rima, *итал.* rima, *нем.* Reim
5. **синонимы**
6. **определение** Созвучие (тождественное или сходное сочетание звуков), систематически повторяющееся в определенном месте стихотворной строки (обычно — в конце).
7. **альтернативные определения** Композиционно-звуковой повтор (преимущественно в конце стихов). Звуковой повтор в конце ритмической единицы.
8. **аннотации (статьи)** Словарь литературных терминов; Литературная энциклопедия; Словарь лингвистических терминов; Поэтический словарь; Краткая литературная энциклопедия
9. **родовое понятие**
10. **видовые понятия** ассонанс, консонанс, диссонанс
11. **целое (к которому относится термин)**
12. **компоненты (части)**
13. **смежность**
14. **ассоциации** вирши, монорим, рифмоид, свободный стих, стихосложение, строфа, эвфония
15. **дисциплина (рубрика первого уровня)** стиховедение
16. **рубрика** стих
17. **источники информации** *Ахманова О. С.* Словарь лингвистических терминов. М.: Сов. энцикл., 1966; *Гончаров Б.* Рифма // Словарь литературоведческих терминов / Ред.-сост. Л. И. Тимофеев и С. В. Тураев. — М.: Просвещение, 1974. — С. 324—326; *Зунделович Я.* Рифма // Литературная энциклопедия: Словарь литературных терминов: В 2-х т. / Под ред. Н. Бродского, А. Лаврецкого, Э. Лунина и др. — М.; Л.: Изд-во Л. Д. Френкель, 1925. — Т. 2: П—Я. — Стб. 717—722; *Квятковский А. П.* Рифма // Квятковский А. П. Поэтический словарь. — М.: Сов. Энцикл., 1966. — С. 248—249; Тимофеев Л. Рифма // Литературная энциклопедия: В 11 т. — М.: ОГИЗ РСФСР, Гос. ин-т. “Сов. Энцикл.”, 1935. — Т. 9. — Стб. 704—708; *Холшевников В. Е.* Рифма // Краткая литературная энциклопедия. — М.: Сов. энцикл., 1971. — Т. 6: Присказка — “Советская Россия”. — С. 306—309.

18. **дополнительные источники информации** *Холшевников В. Е.* Рифма // Большая советская энциклопедия. — 3-е изд. — М.: Сов. энцикл., 1975. — [Т.] 22: Ремень — Сафи...

1. **термин** канцона
2. **варианты написания** кансона
3. **этимология** *итал.* canzone ‘песня’
4. **иноязычные эквиваленты** *итал.* canzone, *франц.* chanson, *прованс.* cansó
5. **синонимы**
6. **определение** Жанр средневековой и ренессансной лирики — стихотворение (преимущественно любовное), содержащее произвольное количество строф с единой схемой рифмовки и особую концевую строфу (торнаду).
7. **альтернативные определения**
8. **аннотации (статьи)** Литературная энциклопедия; Словарь литературных терминов; Поэтический словарь
9. **родовое понятие** лирика
10. **видовые понятия** канцонетта
11. **целое (к которому относится термин)**
12. **компоненты (части)** строфа, торнада
13. **смежность** трубадуры
14. **ассоциации** dolce stil nuovo
15. **дисциплина (рубрика первого уровня)** поэтика
16. **рубрика** теоретическая поэтика : роды, виды, жанры
17. **источники информации** *Квятковский А. П.* Канцона // Квятковский А. П. Поэтический словарь. — М.: Сов. энцикл., 1966. — С. 130; *Никонов В.* Канцона // Словарь литературоведческих терминов / Ред.-сост. Л. И. Тимофеев и С. В. Тураев. — М.: Просвещение, 1974. — С. 121; *Рукавишников И. С.* Канцона // Литературная энциклопедия: Словарь литературных терминов: В 2-х т. / Под ред. Н. Бродского, А. Лаврецкого, Э. Лунина и др. — М.; Л.: Изд-во Л. Д. Френкель, 1925. — Т. 1: А—П. — Стб. 344—345. — Подп. И. Р.; *Сергиевский М.* Канцона // Литературная энциклопедия: В 11 т. — [М.]: Изд-во Ком. Акад., 1931. — Т. 5. — Стб. 121.

18. **дополнительные источники информации** *Зверев Г. И.* Канцона // Краткая литературная энциклопедия. — М.: Сов. энцикл., 1966. — Т. 3: Иаков — Лакснесс. — Стб. 375; Канцона // Большая советская энциклопедия. — 3-е изд. — М.: Сов. энцикл., 1973. — [Т.] 11: Италия — Кваркуш.

1. **термин** варваризм
2. **варианты написания**
3. **этимология** от *лат.* barbarus ‘чужеземный’
4. **иноязычные эквиваленты** *англ.* barbarism, *фр.* barbarisme, *исп.* barbarismo, *нем.* Barbarismus
5. **синонимы**
6. **определение** Слово, заимствованное из иностранного языка
7. **альтернативные определения** Слово, элемент слова или оборот речи, заимствованные из иностранного языка
8. **аннотации (статьи)** Словарь литературных терминов; Словарь лингвистических терминов; Словарь литературоведческих терминов
9. **родовое понятие** заимствование, заимствованная лексика
10. **видовые понятия** англицизм, галлицизм, гебраизм, германизм, грецизм, латинизм, полонизм
11. **целое (к которому относится термин)** лексика
12. **компоненты (части)**
13. **смежность**
14. **ассоциации** неологизм
15. **дисциплина (рубрика первого уровня)** стилистика
16. **рубрика** лексика и фразеология
17. **источники информации** *Ахманова О. С.* Словарь лингвистических терминов. М.: Сов. энцикл., 1966; Варваризм // Литературная энциклопедия: В 11 т. — [М.]: Изд-во Ком. Акад., 1929. — Т. 2. — Стб. 104—105; *Зунделович Я.* Варваризмы // Литературная энциклопедия: Словарь литературных терминов: В 2-х т. / Под ред. Н. Бродского, А. Лаврецкого, Э. Лунина и др. — М.; Л.: Изд-во Л. Д. Френкель, 1925. — Т. 1: А—П. — Стб. 129—130; *Трофимов И.* Варваризм // Словарь литературоведческих терминов / Ред.-сост. Л. И. Тимофеев и С. В. Тураев. — М.: Просвещение, 1974. — С. 36.
18. **дополнительные источники информации**

1. **термин** антанакласис
2. **варианты написания** антанаклазис
3. **этимология** *греч.* ἀντανάκλασις
4. **иноязычные эквиваленты** *греч.* ἀντανάκλασις, *лат.* traductio
5. **синонимы** антанакласа, дилогия
6. **определение** Повторение слова в разных значениях.
7. **альтернативные определения** Повторение слова в противоположных значениях.
8. **аннотации (статьи)** Литературная энциклопедия
9. **родовое понятие** фигура
10. **видовые понятия**
11. **целое (к которому относится термин)**
12. **компоненты (части)**
13. **смежность**
14. **ассоциации**
15. **дисциплина (рубрика первого уровня)** риторика
16. **рубрика** тропы и фигуры
17. **источники информации** Антанаклазис // Литературная энциклопедия: В 11 т. — [М.]: Изд-во Ком. Акад., 1929 (переизд. 1930). — Т. 1. — Стб. 679; ; Шор Р. Фигуры // Литературная энциклопедия: В 11 т. — [М.]: Худож. лит., 1939. — Т. 11. — Стб. 710—713. — Подп. R. S.
18. **дополнительные источники информации** *Дворецкий И. Х.* Древнегреческо-русский словарь. — М., 1958. — Т. I: А — Λ.

Thesaurus as a Poetological Tool

Boykov V.N., Zakharov V.E., Pilshchikov I.A., Sysoev T.M.

Keywords: thesaurus, rubricator, terminology, nomenclature, relations, poetology, poetics

This article discusses the basic principles of the poetological thesaurus embracing all the branches of learning related to comprehensive theoretical and historical study of poetry. This thesaurus is conceived of as a subject-oriented reference system, as an information search tool, and an analytical instrument. The proposed conceptual and engineering design is evaluated from the standpoint of modern standards of thesaurus implementation in information retrieval systems.

Сведения об авторах: Бойков Владимир Николаевич,
Институт космических исследований РАН, консультант;
Захаров Владимир Евгеньевич,
Физический институт им. П.Н.Лебедева РАН, заведующий сектором математической физики;
Пильщиков Игорь Алексеевич,
Институт мировой культуры МГУ им. М.В. Ломоносова, ведущий научный сотрудник;
Сысоев Тимофей Михайлович,
Межведомственный суперкомпьютерный центр РАН, старший научный сотрудник.