Модел. и анализ информ. систем. Т. **20**, № **4** (2013) 125–135

УДК 025.4.06

Тезаурус по поэтологии как инструмент для информационного поиска и коллекции знаний[1]

Бойков В. Н., Захаров В. Е., Каряева М. С., Соколов В. А.

Институт космических исследований РАН,
117997 Россия, Москва, Профсоюзная, 84/32
Физический институт им. П.Н. Лебедева РАН,
119991, Россия, Москва, Ленинский проспект, 53
Ярославский государственный университет им. П. Г. Демидова,
150000 Россия, Ярославль, Советская, 14

e-mail: boykov_bh@bk.ru; zakharov@math.arizona.edu;
mari.karyaeva@gmail.com; valery-sokolov@yandex.ru

получена 16 октября 2013

Ключевые слова: тезаурус, рубрикатор, лингвистические онтологии, терминология, отношения, поэтология, стиховедение

Рассмотрены основные подходы по конструктивному созданию открытого сетевого ресурса «Предметно-специфицированного тезауруса по поэтологии», который является одним из уровней информационно-аналитической системы русской поэзии (ИАС РП). Под поэтологией будем понимать группу дисциплин, ориентированную на всестороннее теоретическое и историческое изучение поэзии. ИАС РП будет использоваться в качестве инструмента для широкого спектра исследований, позволяющего определять характерные признаки анализируемых произведений поэзии. Таким образом, тезаурус соответствует базе знаний, из которой будут заимствоваться исходные данные для обучения системы. Описан подход по формированию базы знаний. Тезаурус представляет собой веб-ресурс, включающий предметно-ориентированный справочник, информационно-поисковый инструмент и инструмент для дальнейших аналитических исследований. Детально рассмотрена проработка терминологического словника, состоящего из 3 тысяч терминов, и комплекса семантических полей. На основании этого представлен rdf-граф предметно-специфицированного тезауруса по поэтологии, содержащий 9 типов объектов и различные виды отношений между ними. Для реализации ресурса применяются Wiki-технологии, что дает возможность хранить данные в форматах Semantic Web.

[1]Работа поддержана Российским фондом фундаментальных исследований, грант № 13-06-00448.

1. Введение

Структурирование накопленной информации в различных предметных областях в настоящее время является первостепенной задачей. За последние несколько десятилетий появились алгоритмы, принципы обработки и хранения больших данных, что дает немаловажный результат для организации баз знаний. Однако для большинства областей наук, которые содержат многоуровневую терминологию, трудно найти правильный подход к формированию баз знаний, тем более сделать единый алгоритм для всех областей знаний не представляется возможным. Безусловно, изменяющиеся концепции определенной предметной области приводят к изменениям в семантике терминологической базы. Вследствие этого, необходимо создание такой системы, которая без проблем имела бы возможность вносить терминологические корректировки. Кроме того, создание базы знаний в предметной области обычно ведется методом проб и ошибок, поэтому возникает необходимость в автоматизированном инструментарии для создания системы. В настоящее время характерно использование в качестве баз знаний таких ресурсов, как предметно-специфицированный тезаурус (ПСТ) [1]. Под ПСТ понимается систематизированная совокупность концептов – терминов и соответствующих им определений понятий, описывающих данную предметную область, с указанием семантических отношений и связей между ними.

Существует достаточно много ресурсов, направленных на решение специфицированных задач. В частности тезаурус WordNet [2], который представлен на многих языках, предназначен для структурирования частей речи. Особенностью ресурса служит отображение слов со схожим значением в виде синонимических рядов. Кроме того, популярный тезаурус Mesh [3], индексирующий статьи по медицине, является уникальным систематизированным ресурсом с точки зрения организации знаний.

Построение ПСТ по поэтологии является схожей задачей, с точки зрения представления семантической обработки информации и использования информационного поиска. Под поэтологией мы будем понимать группу дисциплин, ориентированных на всестороннее теоретическое и историческое изучение поэзии. Предпосылки по созданию тезауруса изложены в работе [4], где собственно сам тезаурус является лишь частью первого блока информационно-аналитической системы русской поэзии (ИАС РП). Структура системы представлена на рисунке 1, она содержит 2 уровня, второй блок будет включать в себя программно-алгоритмический комплекс, исходные данные для которого будут заимствоваться из ПСТ. Таким образом, конечным продуктом будет система, включающая в себя не только формализованную систему понятий в качестве ПСТ, но и инструмент для широкого спектра исследований, позволяющий определять характерные признаки анализируемых произведений поэзии.

2. Разработка терминологического словника

Тезаурус по поэтологии разрабатывается совместно со специалистами в предметной области. На данный момент терминологический словник насчитывает порядка 3000 терминов. Процесс составления словника проходил в несколько этапов: на первом этапе эксперты предметной области провели ручную обработку документов, связан-

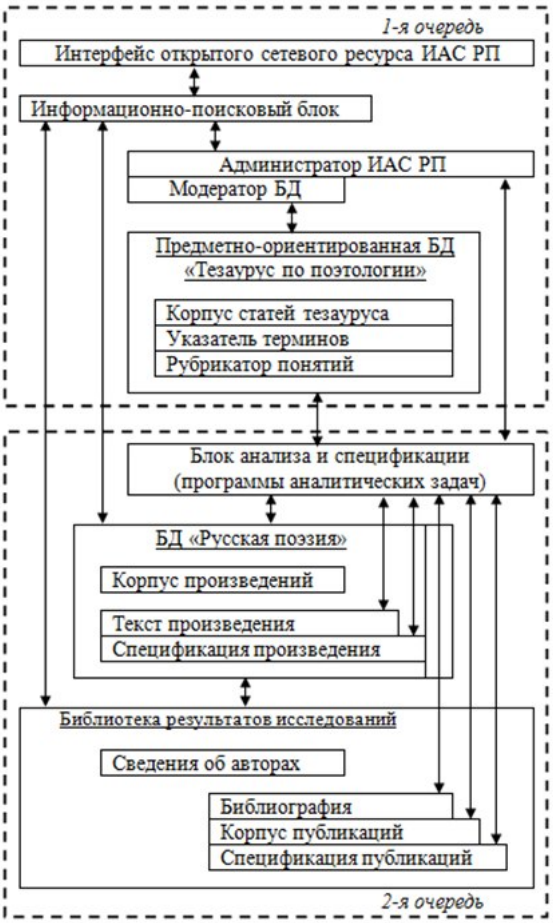

Рис. 1. Структура ИАС РП

ных с предметной областью, и составили единый список терминов. На следующем этапе была составлена классификация верхнего уровня, представляющая подобласти знания, соответствующие дисциплинам поэтологии:

1. Стиховедение;
2. Стиховедение;
3. Метрика;
4. Явления начала и конца стиха (строки);
5. Ритмика;
6. Строфика;
7. Рифмика;
8. Лингвистика стиха;
9. Проза (в отличие от стиха);

10. Стилистика;
11. Поэтика;
12. Риторика;
13. История литературы;
14. Переводоведение и литературная компаративистика;
15. Текстология;
16. Герменевтика;
17. Теоретические школы и направления;
18. Логика и методология науки.

Здесь рубрика «Стиховедение», как наиболее формализованная дисциплина поэтологии, представлена подрубриками 2-го и 3-го уровней. Третий этап составления словника состоял в сведении всех терминов в представленную структуру, которая является основой для создания иерархии. Очевидно, что формальной моделью тезауруса является связный семантический граф, узлами которого служат понятия. В связи с большим объемом терминологического словника понятна целесообразность решения о создании тезауруса как открытого сетевого ресурса с полноценным доступом к созданию и редактированию корпуса терминологических статей.

Не стоит забывать о том, что процесс организации рубрикатора может выявить пробелы, которые могут привести к добавлению новой терминологии, выявлению потребности в новых уровнях в иерархии или, наоборот, возникнет необходимость объединить термины, которые ранее не были признаны идентичными, с точки зрения предметной области.

3. Структура статьи

Каждая терминологическая статья тезауруса (ТСТ) состоит из полей для заполнения, которые делятся на две группы. Первая группа содержит различные атрибуты заглавного поля (понятия) «термин». Вторая – представляет различные, в т.ч. отражаемые в рубрикации, отношения данного «термина» с другими терминами тезауруса. Структура ТСТ в общем определяет логико-семантическую модель «Тезауруса по поэтологии» [5].

Первая группа полей делится в свою очередь на две части: одни из них, существенные для аналитических задач, находятся в разной степени отношения эквивалентности с заглавным термином и содержат различные дополнения к его определению, другие представляют собой разного рода отсылки и комментарии.

Вторая группа семантических полей также делится по двум типам отношений между понятиями: иерархическим (род/виды, целое/части, выше/ниже), определяющим рубрикацию тезауруса, и неиерархическим (соподчинение, смежность, ассоциация, комбинативность), устанавливающим перекрестные связи.

1.	Поля, связанные с определением термина	Термин, варианты написания, иноязычные эквиваленты, синонимы, теоретическое (дескриптивное) определение, альтернативное определение, конструктивное определение, схема(формула), рубрика
2.	Поля, содержащие отсылки и комментарии к термину	Дисциплина (рубрика первого уровня), национальная традиция, автор термина, этимология, аннотации (статьи), примеры употребления, источники информации, дополнительные источники информации, авторы статьи
3.	Поля, содержащие иерархические отношения термина с другими терминами тезауруса	Родовое понятие, видовые понятия, целое, части, условная иерархия (выше/ниже)
4.	Поля, содержащие неиерархические отношения термина с другими терминами тезауруса	Соподчинение, смежность, ассоциации, комбинативность

Таблица 1. Семантические поля ТСТ

Всего же в ТСТ позиционируется 27 полей, и этот список открыт: в ходе постановки и решения аналитических задач может понадобиться позиционирование новых атрибутов и отношений.

4. Отношения между терминами в ПСТ

Единицей тезауруса служит термин, состоящий из слова или словосочетания (терминологического понятия). Отношения в тезаурусе строятся по наличию заполненных семантических полей. Между терминами всегда присутствуют отношения, поскольку существуют обязательные поля для заполнения, например, «синонимы», «дисциплина», «рубрика». Ниже представлены все существующие отношения между терминами в тезаурусе.

1. Отношение синонимии (характеризуется с помощью поля «Синонимы»)
2. Иерархические отношения:
3. Часть-Целое (поля «Целое», «Части»)
4. Выше-Ниже (поля «Ниже», а также «Дисциплина» и «Рубрика»)
5. Родо-видовые отношения (поля «Родовое понятие», «Видовые понятия»)
6. Неиерархические отношения:
7. Отношение соподчинения (поле «Соподчинение») – отношение между видами одного рода, в том числе антонимия терминов.

8. Отношение смежности (поле «Смежность») – фиксация метонимических (переносных) отношений термина.

9. Отношение ассоциации (поле «Ассоциации») – все прочие термины, связанные с тезаврируемым термином, отношения с которыми не определяются в полях «Ниже», «Соподчинение», «Смежность».

10. Отношение комбинативности (поле «Комбинативность») – отношение между членами независимых параллельных классификаций родового понятия.

При заполнении полей может возникнуть ситуация отсутствия необходимого синонима (или другого понятия в поле, обязательном для заполнения), тогда в обязательном поле для заполнения указывается синонимичный термин, который автоматически становится новым объектом тезауруса с пустыми полями.

В тезаурусе предусмотрена индексация рубрик. Каждая статья имеет свою метку для рубрикации, которая задается пользователем при заполнении полей при создании термина.

Нижеприведенный алгоритм позволяет провести первоначальный этап автоматической рубрикации и определить положение нового термина в семантическом графе:

1. Если у термина нет ни рода и ни целого и нет ни вида и ни части, то это означает, что термин вне иерархической рубрикации.

2. Если у термина есть род или целое и есть виды или части, то это означает, что термин является промежуточным в иерархической рубрикации.

3. Если у термина есть род или целое и нет ни вида и ни части, то это означает, что термин является конечным в иерархической рубрикации.

4. Если у термина нет ни рода и ни целого и есть виды или части, то это означает, что термин является начальным в иерархической рубрикации.

5. Инструментальный подход

ПСТ представлен в виде открытого сетевого ресурса с использованием технологии Wiki [6], [7] . Под данной технологией понимается набор средств, для создания интернет-ресурса, позволяющий пользователям, не имеющим навыков программирования и знания языка HTML, создавать, редактировать, вносить дополнения в тезаурус. Сетевой ресурс оснащен аутентификацией и авторизацией пользователей. Кроме того, система предусматривает наличие пользователя, обладающими правами администратора, в обязанности которого входит добавление или удаление внесенных изменений в систему. Статистические характеристики системы определяются количеством посещения каждого пользователя, а также сведениями о его активности. Это позволит анализировать текущую систему с точки зрения дальнейших перспектив развития.

Технологии Wiki работают совместно с реляционной базой данных. База данных содержит 54 таблицы, логически разделенные на группы по определенным критериям, в зависимости от их назначения: статистические таблицы, таблицы аутентификации пользователя, таблицы с категориями, страницами, содержанием статей и так далее.

На первом этапе тезаурус будет заполняться вручную специалистами в предметной области, а также всеми желающими. Далее внесенная информация должна быть подтверждена администратором системы.

6. Семантическая структура тезауруса

Тезаурус может хранить данные в форматах семантической паутины и предоставлять возможность импорта/экпорта в RDF [8]. Это позволяет создавать семантические онтологии, с помощью которых становится доступным расширение модели тезауруса и его наполнения в рамках единой структуры. Для создания онтологии [9] тезауруса используется специализированный формат RDF для представления данных, предназначенных для машинной обработки. Процесс выгрузки или загрузки при описании модели данных тезауруса [10] в данном формате становится оптимальным решением. На рисунке 2 показана схема модели ПСТ для представления ее в формате RDF. Модель ПСТ содержит следующие виды объектов:

1. Термин (term)
2. Понятие (concept)
3. Понятие из Библиотеки Авторов (concept of Authors dictionary = concept of AD)
4. Понятие из Библиотеки пользователей (concept of user dictionary = concept of UD)
5. Понятие из Библиотеки традиций (concept of tradition dictionary = concept of TD)
6. Ссылка (link)
7. Комментарий (annotation
8. Комментарий употребления (example)

Под объектом «понятие» понимается слово или словосочетание, не являющееся термином и имеющее сходные варианты написания с термином. Поэтому такие объекты выделяются в отдельный справочник. «Понятие из Библиотеки Авторов», «Понятие из Библиотеки пользователей», «Понятие из Библиотеки традиций» аналогично являются обособленными частями, каждая из представленных частей несет в себе набор данных об авторах термина, пользователях, которые заполнили статью и о национальной терминологической традиции соответственно. Объект «ссылка»

представляет гиперссылку на источник, поэтому было бы разумнее хранить ее отдельно от всех объектов. «Комментарий» является объектом общей неструктурированной информации. «Комментарий употребления» – наиболее востребованный объект для следующих уровней системы, поэтому выделения его из объекта «Комментарий» позволит ИАС РП обучаться уже на готовых корпусах, состоящих из содержимого объекта. Такой выбор объектов позволяет не только коллекционировать знания, но и проводить дальнейший анализ всей информационно-аналитической системы, где тезаурус является только частью. В модели ПСТ предусмотрены следующие виды отношений между объектами. На рисунке 2 связи, которые являются обязательными, изображены в виде черной стрелки, необязательные отношения – пунктирной синей линией со стрелкой.

1. hasSynonym – отношение синонимии
2. hasAltDef – альтернативное определение
3. hasEthymology – этимология
4. hasSchema – схема (формула)
5. HasHLevel – рубрика
6. sourseOfInf – источник информации
7. hasTradition – национальная традиция
8. hasAdjacency – смежность
9. illustratedBy – примеры употребления
10. hasAuthor – автор термина
11. hasSection – дисциплина
12. hasExtraInf – дополнительные источники информации
13. hasLink – аннотации статьи
14. narrower – отношение ниже
15. hasParallel – комбинативность
16. writtenBy – авторы статьи
17. hasSpelling – варианты написания
18. relatedWith – отношение ассоциации
19. hasHierarchy – соподчинение
20. isPartOf – отношение Целое-Часть
21. hasTypeOf – отношение Выше-Ниже

22. hasThDef – теоретическое (дескриптивное) определение

23. hasConDef – конструктивное определение

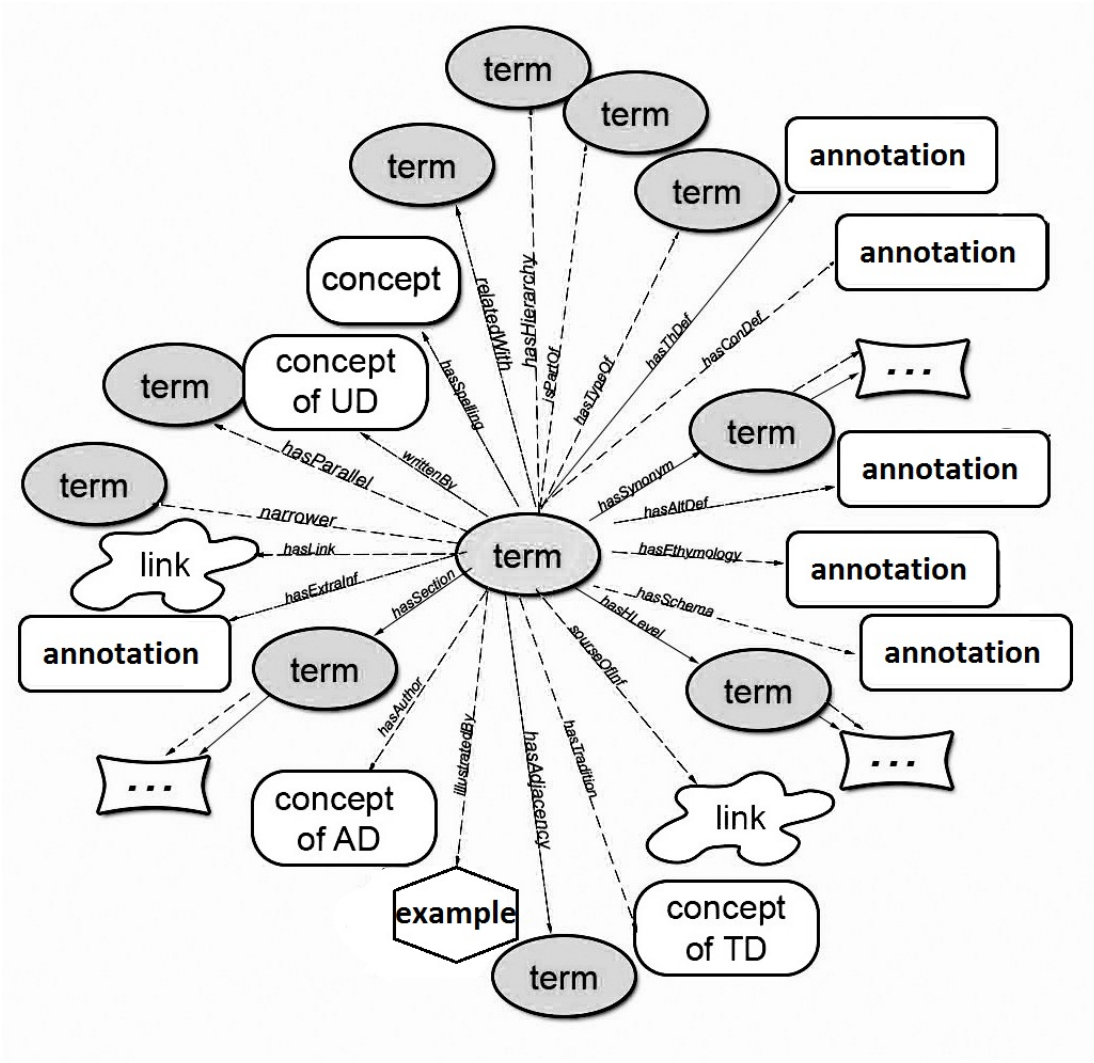

Рис. 2. Схема модели ПСТ для представления в формате RDF

7. Заключение

В работе представлены подходы к разработке ПСТ по поэтологии как открытого сетевого ресурса с использованием Wiki-технологий. Приведена структура тезауруса с описанием полей, семантических отношений, а также схема онтологии для представления в формате RDF.

Список литературы

1. *Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: Изд-во МГУ, 2011.*

2. *Fellbaum C. WordNet: An Electronic Lexical Database. MIT Press, Cambrige, 1998*

3. *Тезаурус MeSH (Medical Subject Headings). URL: http: //www.nlm.nih.gov/mesh/*

4. *Бойков В.Н., Захаров В.Е., Пильщиков И.А., Сысоев Т.М. Тезаурус как инструмент поэтологии // Моделирование и анализ информационных систем. 2010. Т. 17, № 1. С. 5–24. (Boykov V.N., Zakharov V.E., Pilshchikov I.A., Sysoev T.M. Thesaurus as a Poetological Tool // Modeling and analysis of information systems. 2010. V. 17, No 1. P. 5–24 [in Russian]).*

5. *Бойков В. Н., Пильщиков И. А. Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы // Интернет и современное общество: Сборник научных статей: Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013), Санкт-Петербург, 9–11 октября 2013 г. СПб.: НИУ ИТМО, 2013. С. 273–279. (URL: http://conf.infosoc.ru [in Russian]).*

6. *Matthew K.I., Felvegi E., Callaway R.A. // Journal of Research on Technology in Education. 2009. V. 42, Issue 1. ISSN 1539–1523.*

7. *Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А. Предметно-ориентированный тезаурус в открытой информационно-аналитической системе // RCDL. 2013. P. 70–76.*

8. *ANSI/NISO Z39.19–2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Resource Description Framework (RDF). URL: www.w3.org/RDF/*

9. *OWL Web Ontology Language Semantics and Abstract Syntax. URL: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/*

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

Domain-Specific Thesaurus as a Tool for Information Retrieval and Collection of Knowledge

Vladimir N. Boikov, Vladimir E. Zakharov, Mariya S. Karyaeva, Valery A. Sokolov

Space Research Institute of the Russian Academy of Sciences,
Profsoyuznaya Str, 84/32, Moscow, 117997, Russia
P.N. Lebedev Physical Institute of the Russian Academy of Sciences,
Leninskiy pr., 53, Moscow, 119991, Russia
P.G. Demidov Yaroslavl State University,
Sovetskaya str., 14, Yaroslavl, 150000, Russia

Keywords: thesaurus, categories, linguistic ontology, terminology, relationships, poetics, prosody

This paper reports basic approaches to constructive creation of an open resource named "Domain-specified thesaurus of poetics", which is one of the levels of an information-analytical system of the Russian poetry (IAS RP). The poetics is a group of disciplines focused on a comprehensive theoretical and historical study of poetry. IAS RP will be used as a tool for a wide range of studies allowing to determine the characteristic features of the analyzed works of poetry. Consequently, the thesaurus is the knowledge base from which one can borrow input data for training the system. The aim of our research requires a specific approach to formating the knowledge base. Thesaurus is a web-based resource which includes a domain-specific directory, information retrieval tools and tools for further analyzes. The study of glossary consisting of three thousand terms and a set of semantic fields is reviewed in this paper. Rdf-graph of the domain-specified thesaurus of poetics is presented, containing 9 types of objects and different kinds of relationships among them. Wiki-tecnologies are used for implementing a resource which allows to store data in Semantic Web formats.

Сведения об авторах:
Бойков Владимир Николаевич,
Институт космических исследований РАН, консультант
Захаров Владимир Евгеньевич,
Физический институт им. П.Н. Лебедева РАН,
доктор физико-математических наук, академик РАН,
главный научный сотрудник
Каряева Мария Сергеевна,
Ярославский государственный университет им. П.Г. Демидова, магистрант
Соколов Валерий Анатольевич,
Ярославский государственный университет им. П.Г. Демидова,
доктор физико-математических наук, профессор,
зав. кафедрой теоретической информатики