

# Предметно-ориентированный тезаурус в открытой информационно-аналитической системе

Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А.  
Институт космических исследований РАН Ярославский государственный университет  
Москва Ярославль  
[boykov\\_bh@bk.ru](mailto:boykov_bh@bk.ru); [vezakhar@mx.iki.rssi.ru](mailto:vezakhar@mx.iki.rssi.ru); [mari.s.ka@mail.ru](mailto:mari.s.ka@mail.ru); [valery-sokolov@yandex.ru](mailto:valery-sokolov@yandex.ru)

## Аннотация

В работе описывается программная оболочка открытого сетевого ресурса «Тезаурус по поэтологии» в составе Информационно-аналитической системы русской поэзии и предложены технические решения и инструментальные средства для создания этого тезауруса. В результате по предложенной двухуровневой структуре информационно-аналитической системы, состоящей из двух частей, реализован верхний уровень и разработан прототип указанного сетевого ресурса.

## 1. Введение

Рассматривается предметно-ориентированная БД – «Тезаурус по поэтологии», концептуальная модель которого представлена в работе [3] «Тезаурус как инструмент поэтологии». Данный тезаурус должен стать основой «Информационно-аналитической системы русской поэзии» (ИАС РП). Предполагается поочередное создание двух частей названной системы, структура которой представлена на Рис.1.

Верхний уровень (1-я очередь) содержит «Тезаурус по поэтологии», обеспечивающий отслеживаемое модератором метаописание такой предметной области как поэтология, т.е. филологический комплекс стиховедения, поэтики и смежных с ними дисциплин, данные которых необходимы для всестороннего изучения стиха и поэзии. Там же содержится пользовательский комплекс в составе интерфейса и информационно-поискового блока, обеспечивающий открытость системы для получения информации и внесения изменений и дополнений в тезаурус.

Предполагается, что нижний уровень (2-я очередь) содержит:

1) блок анализа с программно-алгоритмическим комплексом, обеспечивающий спецификацию –

выявление (распознавание) существенных (специфицирующих) признаков произведений поэзии;

2) БД поэтических произведений «Русская поэзия»;

3) библиотеку результатов исследования (публикаций) по поэтологии, которые также могут быть объектами исследований.

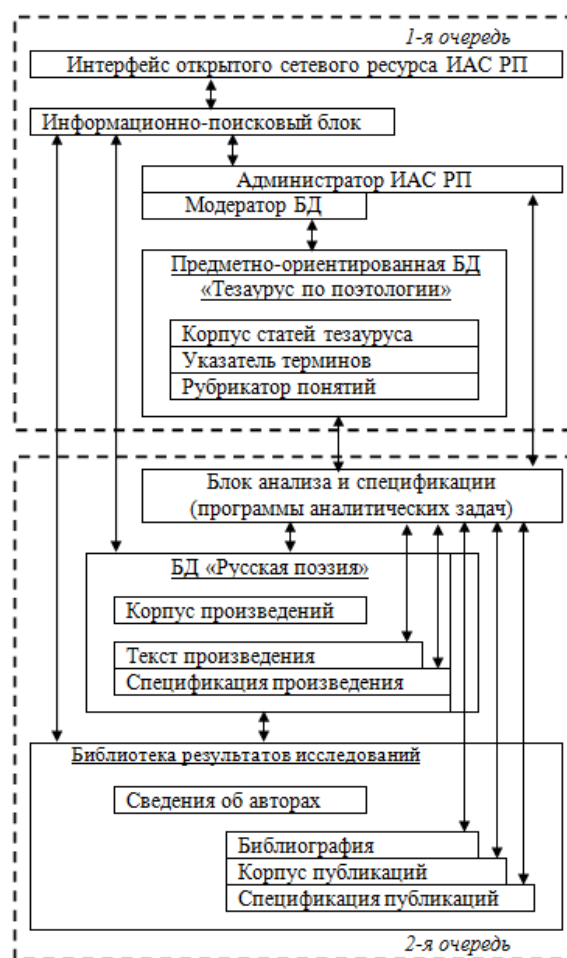


Рис. 1. Структура ИАС РП

Из этой структуры нетрудно видеть, что «Тезаурус по поэтологии» предназначен не только для структурирования предметной области и информационного поиска [9], но главным образом для автоматизированного анализа и спецификации поэтических и поэтологических текстов. Это предъявляет дополнительные требования к информационно-логической модели тезауруса, представленной в разделе 6, которая определяется отношениями между терминами и составом семантических полей в структуре терминологической статьи тезауруса (ТСТ). К примеру, поле «Термин» имеет 7 так или иначе эквивалентных полей («Варианты написания»; «Иноязычные эквиваленты»; «Синонимы»; «Теоретическое (дескриптивное) определение»; «Конструктивное определение»; «Альтернативные определения»; «Схема (формула)»). При этом напрямую связано с автоматизированным анализом текстов необходимость позиционирования таких полей как «Теоретическое (дескриптивное) определение», «Конструктивное определение» и «Схема (формула)», заполнение которых представляет непростую задачу для специалистов. Всего же в ТСТ позиционируются 23 поля, и этот список открыт: в ходе решения задач может понадобиться добавление новых атрибутов и отношений. Основные требования по созданию «Тезауруса по поэтологии» были согласованы со стандартами по созданию информационно-поисковых тезаурусов [5-7, 16-17].

Если учесть, что объем терминологического словника тезауруса на данное время уже насчитывает около 3000 терминов, то становится понятно насколько сложным и трудоемким является процесс создания полного корпуса ТСТ. Ввиду этого очевидна целесообразность доверить решение этой задачи широкому кругу специалистов и осуществить создание «Тезауруса по поэтологии» в качестве открытого сетевого ресурса.

## 2. Обзор инструментальных средств

Тезаурус должен обладать развивающейся структурой, ее содержание, объем и форма должны постоянно корректироваться в соответствии с изменениями, происходящими в обслуживаемой предметной области. В связи с этим должен быть предусмотрен технологический процесс ведения тезауруса, а основным способом доступа к его информационному комплексу 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 системе предполагается возможность загрузки и выгрузки специфицированной части общего массива данных в следующих форматах:

- RDF, который является способом представления данных, предназначенным для машинной обработки. Поскольку он лежит в основе модели данных тезауруса, 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 ТСТ.

- Текстовые файлы. Данный формат наиболее удобен для пользователей, наполняющих тезаурус, поскольку должна быть предусмотрена возможность без особых усилий вносить изменения. Более того, формат текста должен удовлетворять определенным требованиям для дальнейшей машинной обработки.

Понятие онтологии [10,14] является ключевым компонентом Semantic Web, позволяющее специфицировать предметную область с помощью семантической разметки. Использование средств для спецификации онтологий существенно упрощает процесс компьютерной обработки. В качестве инструментальных средств можно отметить следующие: GATE Ontology Editor [13], Prot'eg'e [18], Fluent Editor [15], Викисловарь [4].

Редактор онтологий **Prot'eg'e** является свободным редактором для построения баз знаний, с помощью которого появляется возможность моделирования и экспортирования онтологий в форматы Semantic Web. Стоит принять к сведению, что в редакторе отсутствует возможность просмотра истории внесенных изменений, разделения профиля доступа к данным для групп пользователей и не предусмотрена работа по заполнению системы в параллельном режиме.

**Fluent Editor**, входящий в состав фреймворка Ontorion, является редактором OWL-онтологий [11,19]. Редактор имеет следующие достоинства: быстрота создания онтологий, настройка прав доступа, аудит безопасности, масштабируемость, высокая производительность и безопасность, однако большим минусом является отсутствие русского языка для представления онтологий.

Свободный редактор **GATE Ontology Editor** обеспечивает графическое представление разметки, а также содержит библиотеку классов, реализованных на Java. Сравнительные характеристики (полномочия пользователей, история редактирования и т.д.) схожи с редактором Prot'eg'e.

**Викисловарь**, один из широко распространенных приложений интернет-энциклопедии «Википедия», реализованный на

принципах Wiki. Главным преимуществом средства является широкий выбор расширений, которые позволяют оптимизировать и настраивать систему надлежащим образом. Пользователи имеют разграничение прав на внесение изменений, создание статьи или только режима чтения. Кроме того, все внесенные изменения сохраняются в общую историю. Расширение Semantic MediaWiki (SMW) позволяет импортировать онтологии и использовать в Wiki-системе структуры и связи. А так же помогает искать, организовывать, просматривать содержимое Wiki. При помощи этого расширения тезаурус сможет хранить данные в форматах Семантической паутины и предоставлять возможность импорта/экспорта в RDF [11] и OWL.

В результате сравнения редакторов онтологий, целесообразным оказывается выбор средства с использованием Wiki-технологий (Викисловарь) в качестве прототипа системы коллективного редактирования тезауруса.

Под Wiki-технологиями понимается набор средств построения веб-сайта, позволяющий пользователям самим через веб-интерфейс активно включиться в процесс редактирования его контента: исправления ошибок, добавления новых материалов и т.д. Использование Wiki-технологий для специалистов других областей не требует использования специальных программ, регистрации на сервере и знания языка HTML.

Для разработки самой системы используется движок MediaWiki, который в сочетании с определенным набором расширений превращается в мощный механизм, и с его помощью реализован веб-ресурс с заданными к нему требованиями как к порталу знаний. Кроме того, использование семантического веба позволяет добавлять информацию, пригодную для машинной обработки.

Пилотная версия соответствующей информационно-логической модели системы размещена по адресу [www.wiki-poetics.tw1.ru/](http://www.wiki-poetics.tw1.ru/).

### 3. Движок MediaWiki

Наиболее распространенным инструментальным средством для создания Wiki-систем является популярный Wiki-движок MediaWiki. Именно для этого универсального движка больше всего строится расширений, позволяющих использовать средства семантических технологий. MediaWiki написан на PHP и для хранения данных использует реляционную базу данных (в нашем случае - это база данных MySQL). MediaWiki является свободной программой, и распространяется на условиях Общественной лицензии GNU. Он предоставляет интерфейс работы с базой страниц, разграничение прав доступа к администрированию системы, возможность обработки текста как в собственном формате Wiki-текст, так и в форматах HTML и TeX (для формул), возможность загрузки изображений и других файлов, а также другие возможности. Гибкая

система расширений позволяет пользователям добавлять собственные новые возможности и программные интерфейсы.

### 4. Расширение ядра системы

Одним из главных преимуществ использования движка MediaWiki является возможность использования расширений. Расширение (Extension) – специализированная сборка PHP-кода, которая добавляет новые возможности, а также расширяет функциональность ядра MediaWiki.

Одним из важных расширений для MediaWiki является Semantic MediaWiki (SMW). Это расширение допускает возможность импорта онтологии и позволяет использовать в Wiki-системе структуры и связи, определенные в ней. А так же помогает искать, организовывать, просматривать содержимое Wiki. В то время, как традиционные Wiki содержат только текст, который компьютер не может ни понять, ни обработать, SMW добавляет семантические аннотации, что позволяет Wiki функционировать, используя преимущества баз данных, превращая MediaWiki в полноценное веб-приложение, использующие автоматически обрабатываемые данные со строго определенной семантикой для того, чтобы расширить функциональность Wiki-системы. При помощи этого расширения тезаурус сможет хранить данные в форматах семантической паутины и предоставлять возможность импорта/экспорта в RDF и OWL.

### 5. Концептуальное проектирование базы данных

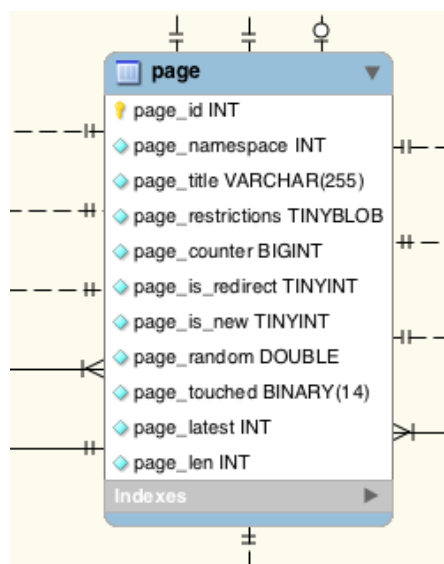


Рис. 2. Таблица «page»

Под концептуальным проектированием понимается построение информационно-логической модели предметной области, то есть модели наиболее высокого уровня абстракции.

Для создания базы данных тезауруса используется 54 таблицы, разделенные на группы по определенным критериям, в зависимости от их назначения. Например, таблицы связанные со статистикой, таблицы с пользователями, категориями, страницами, содержимым страниц и так далее. Ниже приведены примеры наиболее показательных с описанием их полей.

Таблица «page» содержит 10 полей плюс идентификатор страницы.

- \*page\_id – первичный ключ
- \*page\_namespace – специальный идентификатор, использующийся для хранения идентификатора в пространстве имен. Диапазон значений от 0 до 15 для стандартного пространства имен, и от 100 до 2147483647 для пользовательских имен.
- \*page\_title – название страницы
- \*page\_restrictions – ограничения страницы, для защиты видимости страницы от определенных пользователей
- \*page\_counter – счетчик количества просмотров страницы
- \*page\_is\_redirect – существует или нет ссылка на другую страницу
- \*page\_is\_new – только что созданная страница или нет
- \*page\_random – поле необходимое при выборе «случайной страницы»
- \*page\_touched – отметка, обновляется при изменении страницы
- \*page\_latest – внешний ключ документа, равен 0 во время создания страницы
- \*page\_len – несжатая длина текста страницы в байтах

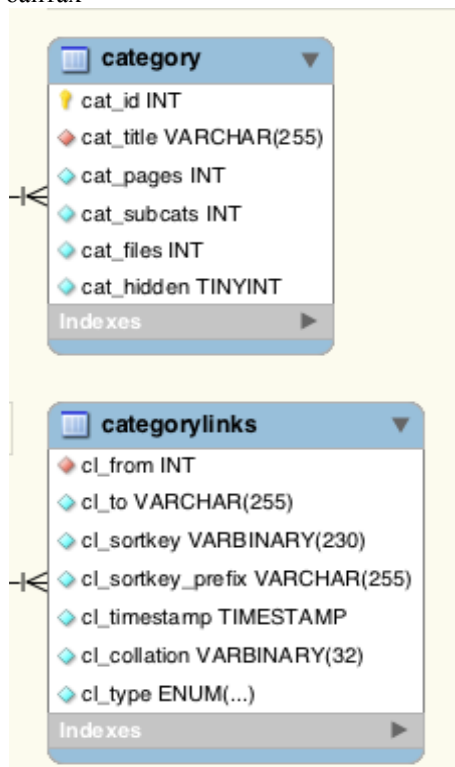


Рис. 3. Таблицы «category» и «categorylinks»

- \*cat\_id – первичный ключ
- \*cat\_title – название категории
- \*cat\_pages – количество страниц в данной категории. Это число включает в себя количество подкатегорий.
- \*cat\_subcats – количество подкатегорий в категории.
- \*cat\_files – количество файлов в категории.
- \*cat\_hidden – зарезервированное поле
- \*cl\_from – сохраняет идентификатор страницы откуда была взята ссылка
- \*cl\_to – хранит название категории
- \*cl\_sortkey – хранит в байтовом коде последовательность отсортированных страниц категории
- \*cl\_timestamp – сохраняет время, когда последний раз обновлялись данные
- \*cl\_sortkey\_prefix – либо пустая строка, либо версия cl\_sortkey созданная администратором
- \*cl\_collation – параметры сортировки
- \*cl\_type – вид статьи (файл, подкатегория или обычная страница)

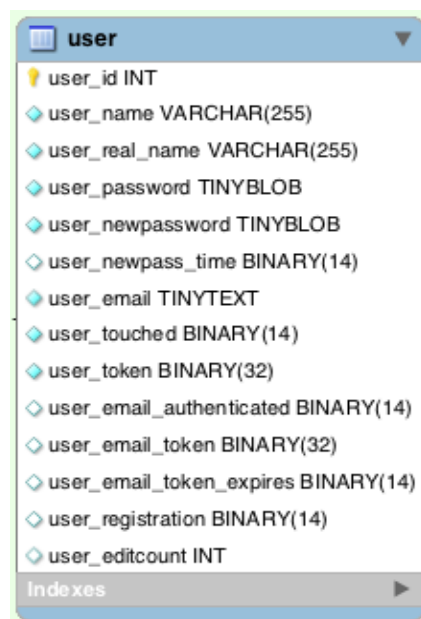


Рис. 4. Таблица «user»

- \*user\_id – первичный ключ, используемый для однозначной идентификации пользователя
- \*user\_name – логин пользователя
- \*user\_real\_name – Настоящее имя пользователя
- \*user\_password – пароль пользователя
- \*user\_newpassword – создается для отправки электронным письмом
- \*user\_newpass\_time – установлен на текущее время. Нужен для установки нового пароля
- \*user\_email – электронная почта пользователя
- \*user\_options
- \*user\_touched – внесенные пользователем изменения

\*user\_token - поле, обеспечивающее возможность сохранить данные пользователя в браузере  
 \*user\_email\_authenticated - метка времени, когда пользователь прошел проверку подлинности по электронной почте  
 \*user\_email\_token - маркер для подтверждения электронной почты  
 \*user\_registration - метка времени, когда пользователь был зарегистрирован  
 \*user\_editcount - счетчик количества редактирований

## 6. Информационное наполнение тезауруса

Заглавная страница сайта включает в себя необходимую справочную информацию о тезаурусе по поэтологии, поле для добавления ТСТ, поисковый блок, а также ссылки на уже созданные ТСТ. Разбивка заглавной страницы состоит из нескольких таблиц, семантических полей поиска и добавления ТСТ, боковая панель с инструментами и меню с регистрационной формой в верхней части экрана.

Для добавления ТСТ необходимо создать семантическую форму для ввода в нее информации, а также семантический шаблон – правило, по которому оформляется страница с данными из формы. На рисунке 5 показана семантическая форма, которая содержит 23 поля, 5 из которых являются обязательными для заполнения. Кроме того, поле «Дисциплина» выбирается из уже первоначально заданных категорий. Форма задана таким образом, чтобы ТСТ автоматически распределялась в соответствующую рубрику [1,2,8] и в дальнейшем отражалась в графе в виде узла.

Рис. 5. Семантическая форма для создания ТСТ

Для создания предупреждений необходимо было внести изменения при помощи Wiki-разметки

непосредственно в саму форму. Исходный код с продемонстрирован ниже и представляет только визуальную часть тезауруса.

```
<includeonly>
<div id="wikiPreview" style="display: none;
padding-bottom: 25px; margin-bottom: 25px; border-
bottom: 1px solid #AAAAAA;"></div>
{{{for template|Статья}}}
{| class="formtable"
! Термин* :
| {{{field|Термин|mandatory}}}
|-
! Автор термина :
| {{{field|Автор термина}}}
|-
! Этимология :
| {{{field|Этимология}}}
|-
! Иноязычные эквиваленты :
| {{{field|Иноязычные эквиваленты}}}
|-
! Синонимы* :
| {{{field|Синонимы}}}
|-
! Теоретическое (дескриптивное)
определение* :
| {{{field|Теоретическое (дескриптивное)
определение|mandatory}}}
|-
! Конструктивное определение* :
| {{{field|Конструктивное
определение|mandatory}}}
|-
! Альтернативные определения :
| {{{field|Альтернативные определения}}}
|-
! Схема :
| {{{field|Схема}}}
|-
! Аннотации (статьи) :
| {{{field|Аннотации (статьи)}}}
|-
! Примеры употребления (цитаты) :
| {{{field|Примеры употребления
(цитаты)}}}
|-
! Родовое понятие :
| {{{field|Родовое понятие}}}
|-
! Видовые понятия :
| {{{field|Видовые понятия}}}
|-
! Целое :
| {{{field|Целое}}}
|-
! Компоненты (части) :
| {{{field|Компоненты (части)}}}
|-
! Смежность :
| {{{field|Смежность}}}
```



```

|-
! Ассоциации :
| {{{field|Ассоциации}}}
|-
! Дисциплина* :
| {{{field|Дисциплина|input
type=combobox|mandatory|property=has
rubric|default=Стихovedение|values=Стихovedение,
Стилистика, Поэтика, Риторика, История
литературы, Переводоведение и литературная
компаративистика, Текстология, Герменевтика,
Теоретические школы и направления, Логика и
методология науки|existing values only}}}
|-
! Рубрика* :
| {{{field|Рубрика|mandatory}}}
|-
! Источники информации :
| {{{field|Источники информации}}}
|-
! Дополнительные источники информации :
| {{{field|Дополнительные источники
информации}}}
}}
{{{end template}}}

***- поля, обязательные для заполнения"

{{{standard input|summary}}}
{{{standard input|minor edit}}} {{{standard
input|watch}}}
{{{standard input|save}}} {{{standard
input|preview}}} {{{standard input|changes}}}
{{{standard input|cancel}}}
</includeonly>

```

Рис.6. Предупреждение об обязательных полях для заполнения в семантической форме

На рисунке 7 показана готовая ТСТ, созданная с использованием формы ввода. Такой способ создания статей является наиболее оптимизированным с точки зрения заполнения тезауруса специалистами предметной области.

Кроме того, предусматривается возможность дальнейшего редактирования ТСТ, внесения изменений и дополнений в уже существующую ТСТ. Созданная ТСТ может включать в себя ссылки на другие ТСТ тезауруса, таким образом, информационно-поисковая система представляет граф, узлами которого являются данные.

Готовая ТСТ должна автоматически размещаться в соответствующей рубрике. На сайте создано дерево категорий, в котором может храниться вся иерархия тезауруса (Рис.8).

Рис.7. ТСТ «Дольник» в тезаурусе

Рис.8. Рубрикация тезауруса

## 7. Заключение

В работе представлена программная реализация предметно-ориентированной БД «Тезаурус по поэтологии».

На сайте [www.wiki-poetics.tw1.ru/](http://www.wiki-poetics.tw1.ru/) приведены примеры заполнения нескольких ТСТ, которые могут быть отредактированы повторно при необходимости только зарегистрированными пользователями, а также администратором. Кроме этого, ему предоставляется возможность отслеживать все изменения ТСТ, при необходимости удалять ненужную информацию и

вносить корректировки по работе сайта. Изменение внешнего вида или внутренних настроек осуществляется тоже администратором, но уже непосредственно на хостинге, где размещен сайт.

Конструктор сайта позволяет изменять программный код сайта, а управление базой данных дает возможность создавать таблицы и доставать информацию из них.

Решение задачи автоматической рубрикации тезауруса является основой для нижнего уровня (2-й очереди) информационно-аналитической системы.

При реализации 2-й очереди системы предстоит сделать выгрузку из базы данных для последующей работы с накопленной информацией. Программно-аналитический комплекс 2-й очереди системы будет предназначаться для решения задач метроритмической спецификации стиха.

На заключительном этапе предполагается разработка программного обеспечения на основе веб-технологий для генерации аналитического комплекса в сетевом пространстве.

## Литература

- [1] Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП 2004 // Российский семинар по Оценке Методов Информационного Поиска.
- [2] Азарова И.В., Синопальникова А.А., Смирн П. Представление устойчивых лексических сочетаний в компьютерном тезаурусе RussNet. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог, 2005.М.: Наука. с. 11-16
- [3] 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.
- [4] Викисловарь // — Режим доступа: <http://ru.wiktionary.org>
- [5] ГОСТ 7.74-96 СИБИД. Информационно-поисковые языки. Термины и определения. Межгосударственный стандарт. Межгосударственный совет по стандартизации, метрологии и сертификации. Минск
- [6] ГОСТ 7.25–2001. СИБИД. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления.
- [7] ГОСТ 7.24-2007. СИБИД. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению
- [8] Клюкин В. Э. Композиция ГЛОТЕОН: глоссарий + тезаурус + онтологии. // НАУКА И ОБРАЗОВАНИЕ: ЭЛЕКТРОННОЕ НАУЧНО-ТЕХНИЧЕСКОЕ ИЗДАНИЕ. - сентябрь 2011г, № 09. Издательство: Московский государственный технический университет им. Н.Э. Баумана (Москва) ISSN: 1994-0408.
- [9] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во МГУ, 2011.
- [10] 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>
- [11] ANSI/NISO Z39.19–2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.Resource Description Framework (RDF) // — Режим доступа: [www.w3.org/RDF/](http://www.w3.org/RDF/)
- [12] 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/](http://www.w3.org/TR/owl-ref/)
- [13] GATE Ontology Editor // — Режим доступа: <http://gate.ac.uk/releases/gate-5.2.1-build3581-ALL/doc/tao/splitch14.html>
- [14] Gruber T.R. The role of common ontology in achieving sharable, reusable knowledge bases // Principles of Knowledge Representation and Reasoning. Proceedings of the Second International Conference. J.A. Allen, R. Fikes, E. Sandewell (Eds.) Morgan Kaufmann, 1991, 601-602.
- [15] Редактор онтологий Fluent Editor // — Режим доступа: <http://www.cognitum.eu/semantics/FluentEditor2/Default.aspx>
- [16] ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri.
- [17] ISO 5964:1985. Guidelines for the establishment and development of multilingual thesauri.
- [18] The Protégé Ontology Editor and Knowledge Acquisition System // — Режим доступа: <http://protege.stanford.edu/>
- [19] W3C OWL Working Group. OWL 2. Web Ontology Language: Document Overview — Режим доступа: [www.w3.org/TR/owl2-overview/](http://www.w3.org/TR/owl2-overview/)

## Domain-Specific Thesaurus as a Part of an Information-Analytical System

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

The work presents an open resource «Thesaurus of Poetics» as the first part of an Information-Analytical System of the Russian Poetry.