



МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(национальный исследовательский университет)»

Кафедра 319 «Системы интеллектуального мониторинга»

Семантически-ориентированный естественно-языковой интерфейс для взаимодействия с Системой взаимосвязанных открытых данных (Linked Open Data)

Автор:
студент группы МЗО-435Б-18
Урубков В.С.
Научный руководитель:
д.т.н. профессор Фомичев В.А.

Москва, 2022

Необходимо разработать семантически-ориентированный естественно-языковой (русскоязычный) интерфейс для взаимодействия с системой взаимосвязанных открытых данных.

Для этого требуется:

- выбрать подход к описанию семантического представления текста на естественном языке;
- разработать алгоритм для реализации преобразования «ЕЯ-запрос → SPARQL-запрос».

Linked Open Data (LOD)

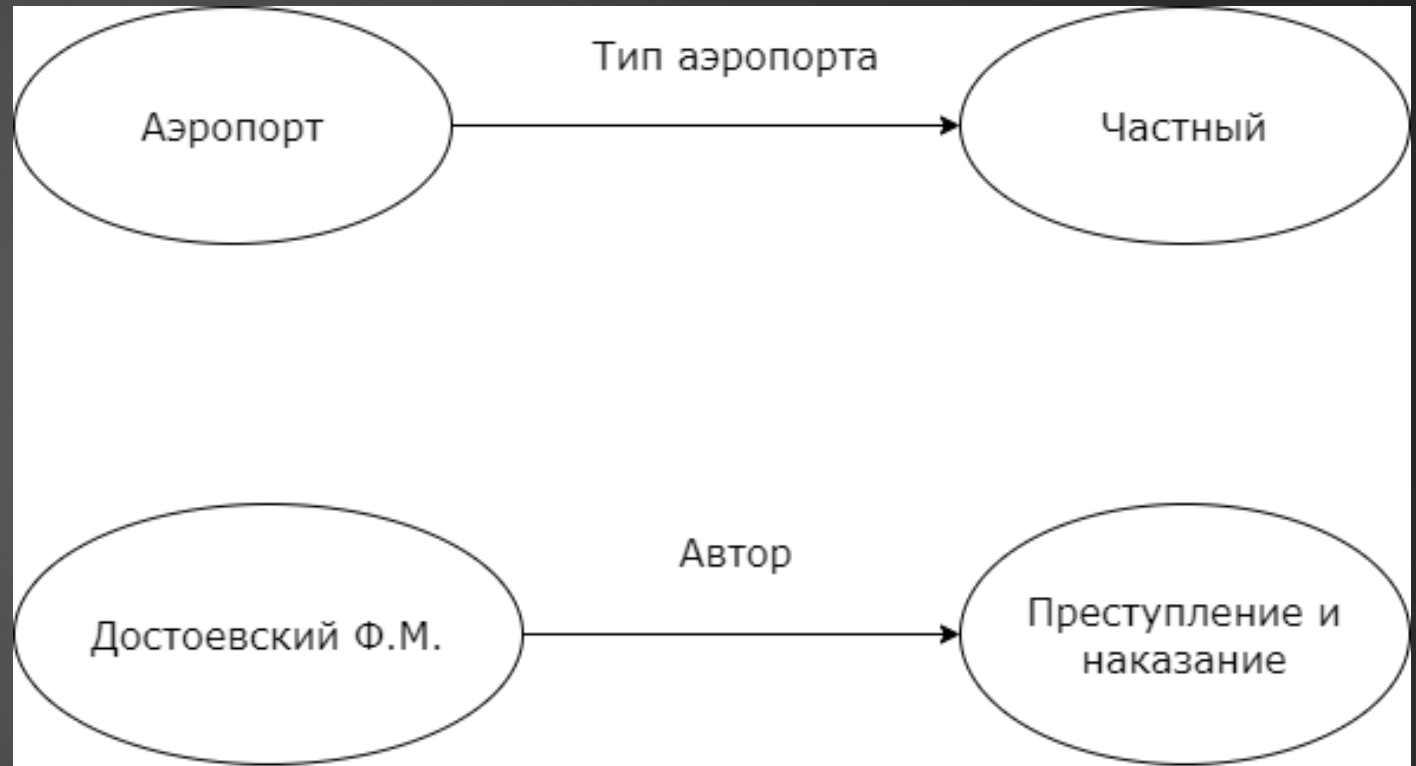
3

Система LOD – огромный размеченный ориентированный граф, состоящий из элементарных графов, представляющих тройки вида (субъект, предикат, объект).



Resource Description Framework – язык для создания распределенных баз знаний (онтологий).

Основной структурой языка RDF являются триплеты – упорядоченные тройки вида (субъект, предикат, объект)



SPARQL

5

SPARQL (рекурсивный акроним SPARQL Protocol and RDF Query Language) – язык запросов к данным, представленным в формате RDF.

```
SELECT DISTINCT ?name
WHERE {
    ?town rdf:type dbo:City.
    ?town dbo:country dbr:Russia.

    ?town rdfs:label ?name.
    FILTER(lang(?name) = "ru").
}
```

name
"Ульяновск"@ru
"Астрахань"@ru
"Бийск"@ru
"Челябинск"@ru

Подходы к описанию семантического представления текстов на естественном языке

Абстрактное
представление
смысла

Теория К-представлений
В.А. Фомичева

Грамматика
Монтегю

Сравнение подходов

7

Критерий	Абстрактное представление смысла	Грамматика Монтегю	Теория К-представлений В.А. Фомичева
Язык текста	Английский	Английский	Русский, Немецкий, Французский, Английский и др.
Типы текстов	Повествовательные предложения	Повествовательные предложения и вопросы	Фразы-высказывания, повествовательные тексты, команды, вопросы
Допустимая структура текста	Отдельное предложение	Отдельное предложение	Связный текст наравне с отдельными предложениями

Лингвистическая база данных

8

Состоит из:

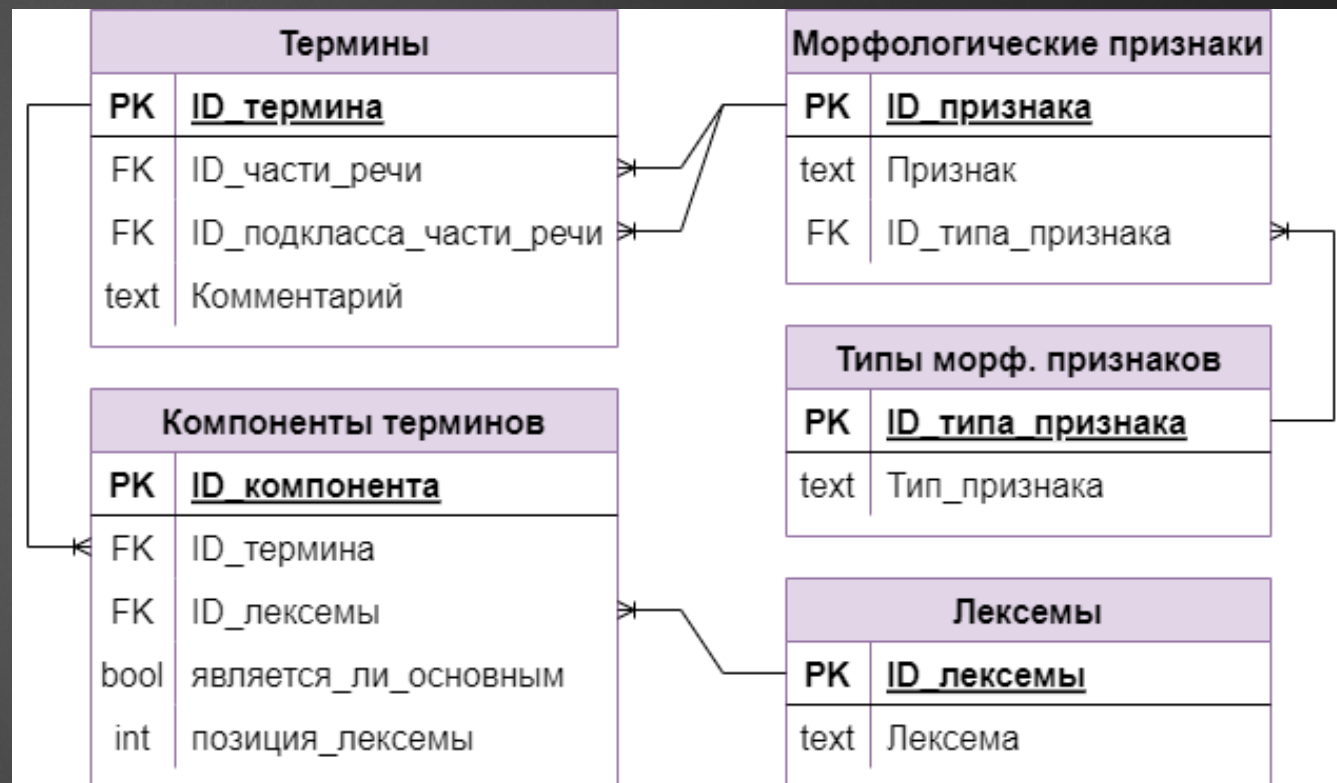
- ▶ Морфологической базы данных
- ▶ Лексико-семантического словаря
- ▶ Словаря предложных фреймов

Морфологическая база данных

9

Содержит информацию о лексемах, терминах, а также набор возможных морфологических признаков.

Для определения морфологических признаков используется нейросетевая библиотека DeepMorphy.



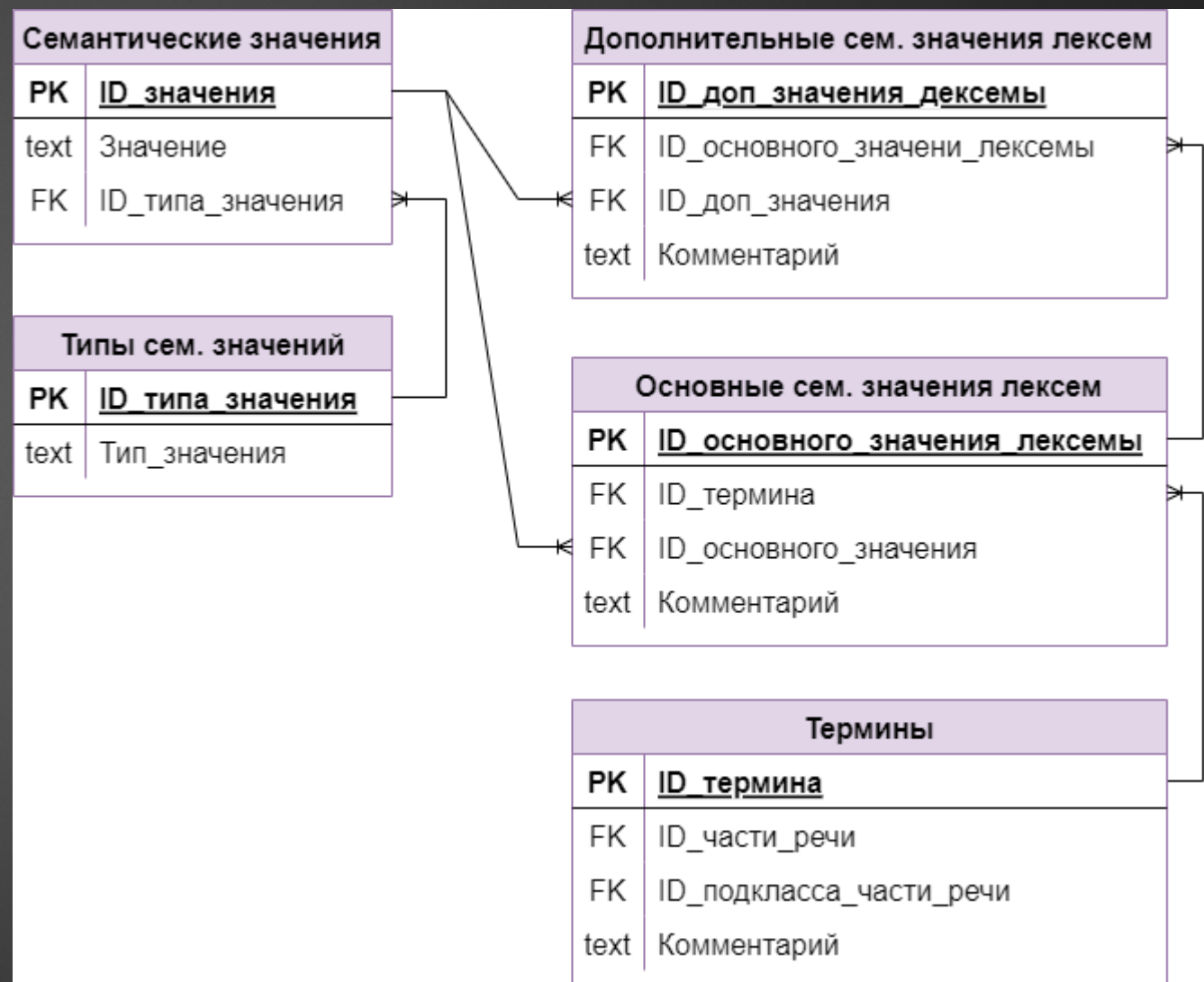
Лексико-семантический словарь

10

Содержит семантические значения лексем и фреймов.

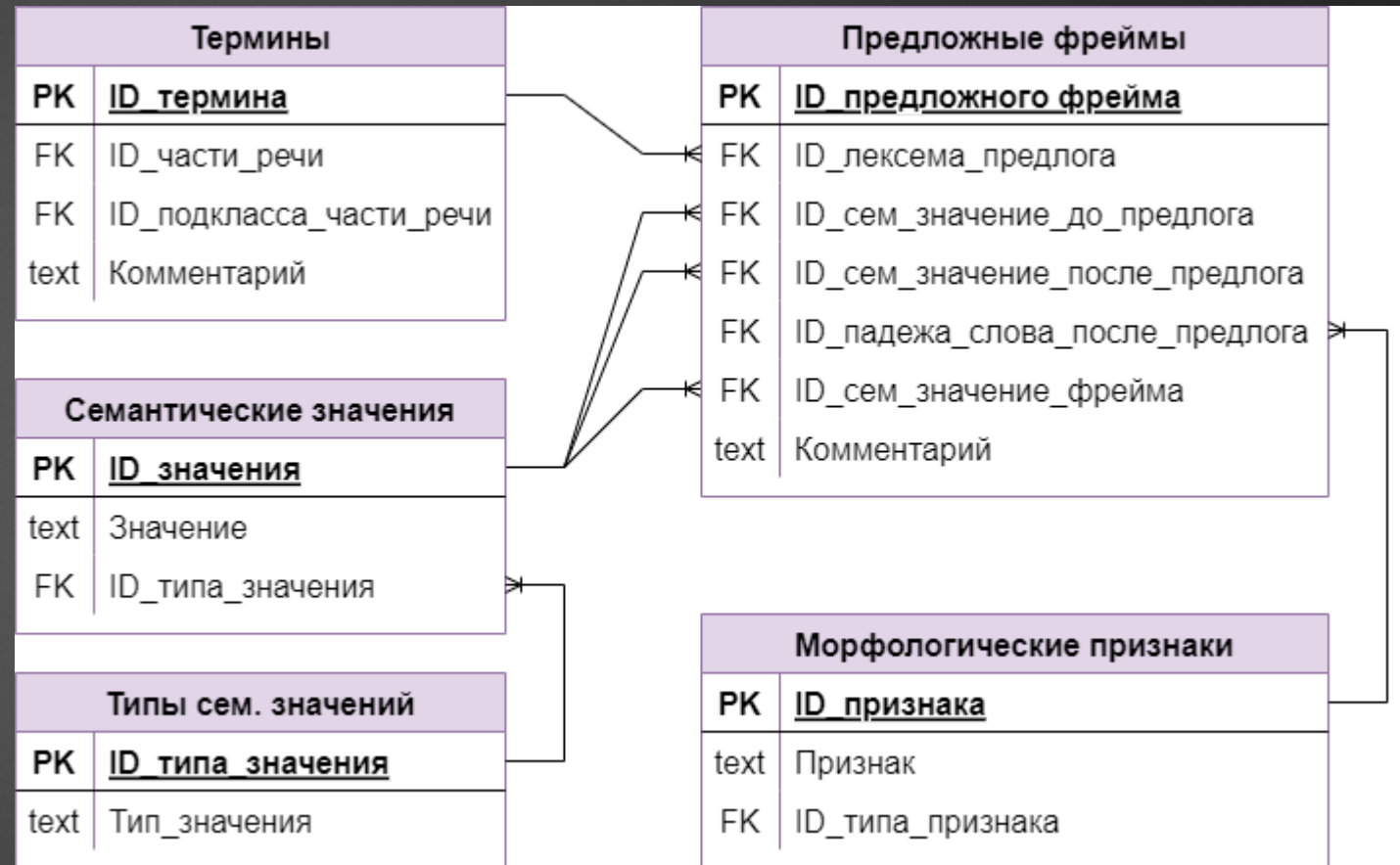
Типы семантических значений:

- Основное значение
- Дополнительное значение
- Значение фрейма



Словарь предложных фреймов

Предложный фрейм описывает смысловые отношения в сочетаниях «Существительное1 + Предлог + Существительное1» и «Существительное1 + Существительное2»



Структура входного запроса на русском языке

12

Фрагмент1 Сущ1 Предлог Фрагмент2 Сущ2 Фрагмент3

- ♦ Фрагмент1 и Фрагмент2 являются либо пустой цепочкой, либо последовательностью прилагательных
- ♦ Сущ1 и Сущ2 – существительные
- ♦ Фрагмент3 является либо пустой цепочкой, либо искусственным именем, либо словосочетанием, определяющим сравнение с числом (например, «меньше 50000» или «не больше 60»)

Примеры запросов

Возможные входные запросы:

- ▶ Планета с самым большим радиусом
- ▶ Одноместные многоцелевые боевые самолёты российского производства
- ▶ Экспериментальные летательные аппараты Китая
- ▶ Широкофюзеляжные самолёты компании Airbus
- ▶ Частные аэропорты Германии
- ▶ Канадские города с населением меньше 50000

Структура семантического представления

14

A (B1, R1, C1) (B2, R2, C2) ... (Bn, Rn, Cn)

- ♦ **A** – обозначение понятия на русском языке (самолёт, автомобиль, компания и т.д.)
- ♦ **B1, B2, ..., Bn** – имена смысловых параметров семантического представления на русском языке
- ♦ **R1, R2, ..., Rn** – имена бинарных отношений на русском языке
- ♦ **C1, C2, ..., Cn** – обозначения значения параметра или второго атрибута отношения на русском языке

Неоднозначность именования в ОНТОЛОГИЯХ

Предикаты, описывающие одно и то же отношение между объектами, могут иметь разные имена даже в рамках одной онтологии.

Город	Предикат
Оттава	population
Москва	populationTotal
Ульяновск	p
Северодвинск	pop2010census

Недостаточная связанность данных

В онтологии YAGO, использующей систему типов Schema:

- ▶ У любых объектов класса Автомобиль отсутствуют содержательные предикаты, хотя информация для них есть
- ▶ Объект класса Город связывается со страной, в которой он располагается, с помощью строки с комментарием типа «Это столица России»

rdfs:comment

"сталица Pacei"@be-tarask

"Столица на Русия"@bg

"capital de Rússia"@ca

"hlavní město Ruské federace"@cs

"Hovedstad i Rusland"@da

Принципы преобразования параметров запросов к LOD

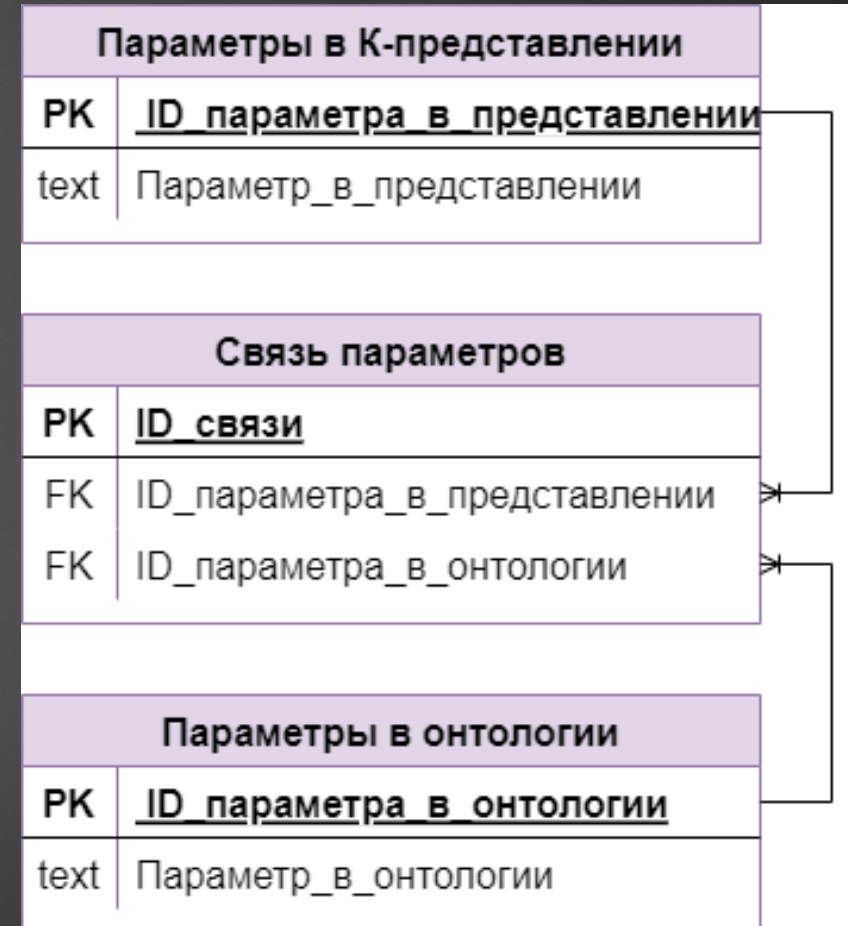
Для обеспечения перевода запроса к LOD на естественном языке в запрос на языке SPARQL необходимо заранее связывать параметры запроса (отношения и некоторые значения) с аналогичными параметрами онтологии.

Отношение	Предикаты
Колич-Жителей	population
	populationTotal
	p
	pop2010census
Россия	dbr:Russia

Компонент разрешения имен

18

Необходим для
связывания параметров
семантического
представления с
параметрами онтологии



Преобразование семантического представления в SPARQL-запрос

SPARQL-запрос условно можно поделить на части следующих типов:

- ▶ Заголовок
- ▶ Тройки равенства
- ▶ Тройки сравнения

Заголовок SPARQL-запроса

20

В заголовке определяется тип искомой сущности на основе понятия **A**, указанного в семантическом представлении входного запроса

Самолёт



```
SELECT DISTINCT ?var1
WHERE {
  VALUES ?var2 {dbo:Aircraft} .
  ?var1 rdf:type ?var2 .
}
```


Тройки равенства

21

(Страна, =, Россия)



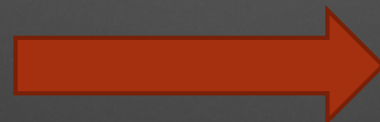
```
VALUES ?p3 {dbo:country} .  
VALUES ?var3 {dbr:Russia} .  
?var1 ?p3 ?var3 .
```

(Экипаж, =, 4)



```
VALUES ?p5 {dbo:Crew}  
.?var1 ?p5 4.
```

(Радиус, =, #макс#)



```
VALUES ?p4 {dbo:radius,  
dbo:meanRadius} .  
?var1 ?p4 ?var4 .  
...  
} ORDER BY DESC (?var4) LIMIT 1
```

Тройка сравнения

22

(Колич-жителей, <, 50000)



```
VALUES ?p4 {dbp:population  
dbp:populationTotal dbp:p  
dbp:pop2010census} .  
?var1 ?p4 ?var4 .  
FILTER (?var4 < 50000) .
```

Средства разработки

23

Для программной реализации приложения были выбраны платформа **.NET** (версия **.NET6**) и язык программирования **C#**. В качестве СУБД выбрана **PostgreSQL**.

Использовались следующие инструменты и технологии:

- ▶ **LINQ to Entity Framework** (взаимодействие с базой данных),
- ▶ **WPF** (оконное приложение),
- ▶ **DeepMorphy** (определение морфологических признаков),
- ▶ **dotNetRdf** (выполнение SPARQL-запросов).

Интерфейс приложения

24

Естественно-языковой интерфейс к LOD

Введите запрос на русском языке

Одноместные многоцелевые боевые самолёты российского производства

Выполнить

Результат выполнения запроса
http://dbpedia.org/resource/Mikoyan_MiG-35
http://dbpedia.org/resource/Sukhoi_Su-35
http://dbpedia.org/resource/Sukhoi_Su-27
http://dbpedia.org/resource/Sukhoi_Su-57

В ходе данной работы:

- ▶ Был разработан и реализован семантически-ориентированный естественно-языковой интерфейс для взаимодействия с Системой взаимосвязанных открытых данных
- ▶ Проведено сравнение подходов к формальному описанию семантической структуры текстов и выбран подход теории К-представлений В.А. Фомичева
- ▶ Разработана и реализована лингвистическая база данных
- ▶ Разработан алгоритм для реализации преобразования «ЕЯ-запрос → SPARQL-запрос»
- ▶ Предложен принцип преобразования параметров запроса, позволяющий преодолевать проблему неоднозначности имен в онтологии