Теория неформальных языков: опыт разработки DSL для запросов к базе знаний

Артём Белоусов

О себе

- В 2023 году окончил бакалавриат ИУ9
- В 2023 году поступил в магистратуру «Системное программирование» ВШЭ
- С осени 2023 года работаю в ИСП РАН
  Разрабатываю:
  - Предметно-ориентированный язык запросов TQL
  - Бэкенд поиска платформы Talisman
- В 2025 году окончил магистратуру

Почему «Системное программирование»?

- Практическая направленность предметов
- Углубление знаний о системном ПО
- Курсы МагоЛего ВШЭ

Поступление в магистратуру

- Единственное вступительное испытание — собеседование
- Программа собеседования хорошо стыкуется с программой ИУ9
- Оценивается не только формальный ответ на вопрос, но и общий кругозор, мотивация

Почему ИСП РАН?

- Легко совмещать с учёбой в магистратуре
- Решение нетипичных задач
- Возможность применить знания с ИУ9 на практике

Устройство в ИСП РАН

- Подал заявку на курсовой проект по разработке языкового сервера из-за увлечения Neovim
- Выполнил тестовое задание на Scala
- Начал создавать языковой инструментарий с нуля на Rust

I семестр

- Верификация программного обеспечения
- Компьютерные сети и телекоммуникации
- Прикладной системный анализ
- Формальные методы программной инженерии

МагоЛего

- Выбор из большого количества курсов разных департаментов ВШЭ
- Количество мест на каждый курс ограничено
- Не всегда получается записаться на интересные курсы

II семестр

- Компиляторные технологии
- Конструирование ядра операционных систем
- Перспективные системы управления базами данных

III семестр

- Компиляторные технологии 2
- Формальные методы программной инженерии. Дополнительные главы
- Теоретическая криптография
- Параллельное программирование

Платформа Talisman и язык TQL

- Talisman — платформа для построения информационно-аналитических систем
- База знаний платформы Talisman представляет из себя граф знаний
- Для поиска объектов в базе знаний используется язык TQL

Зачем здесь вообще DSL?

- Поиск сразу в нескольких СУБД (PostgreSQL, OpenSearch)
- Пользователи платформы аналитики, а не программисты
- Гибкое разграничение прав доступа

Внешние и встраиваемые DSL

Встраивамые DSL:

- Реализуются как библиотеки или макросы
- Пользователю доступен стандартный инструментарий
- Гибкость ограничена хостовым языком

Внешние DSL:

- Реализуются независимо от других языков
- Инструментарий разрабатывается отдельно
- Большая гибкость при разработке

Пример TQL запроса

```
"Семинар META"~2 Концепт.студент(
  Связь>."учится в"(
    Концепт.университет(
      "год основания"=1830, город=Долгопрудный,
субъект=Татарстан)))
```

Вопросы по TQL (1/4)

Чем отличаются запросы?

- `Концепт(имя=Саша Связь>."работает в"(Концепт(имя=яндекс)))`
- `Концепт(имя=яндекс Связь<."работает в"(Концепт(имя=Саша)))`

Вопросы по TQL (2/4)

Эквивалентны ли запросы?

- `Концепт(город=Переславль название="ИПС РАН")`
- `Концепт(город=Переславль) Концепт(название="ИПС РАН")`

Вопросы по TQL (3/4)

Эквивалентны ли запросы?

- `Концепт(город=Переславль,название="ИПС РАН")`
- `Концепт(город=Переславль),Концепт(название="ИПС РАН")`

Вопросы по TQL (4/4)

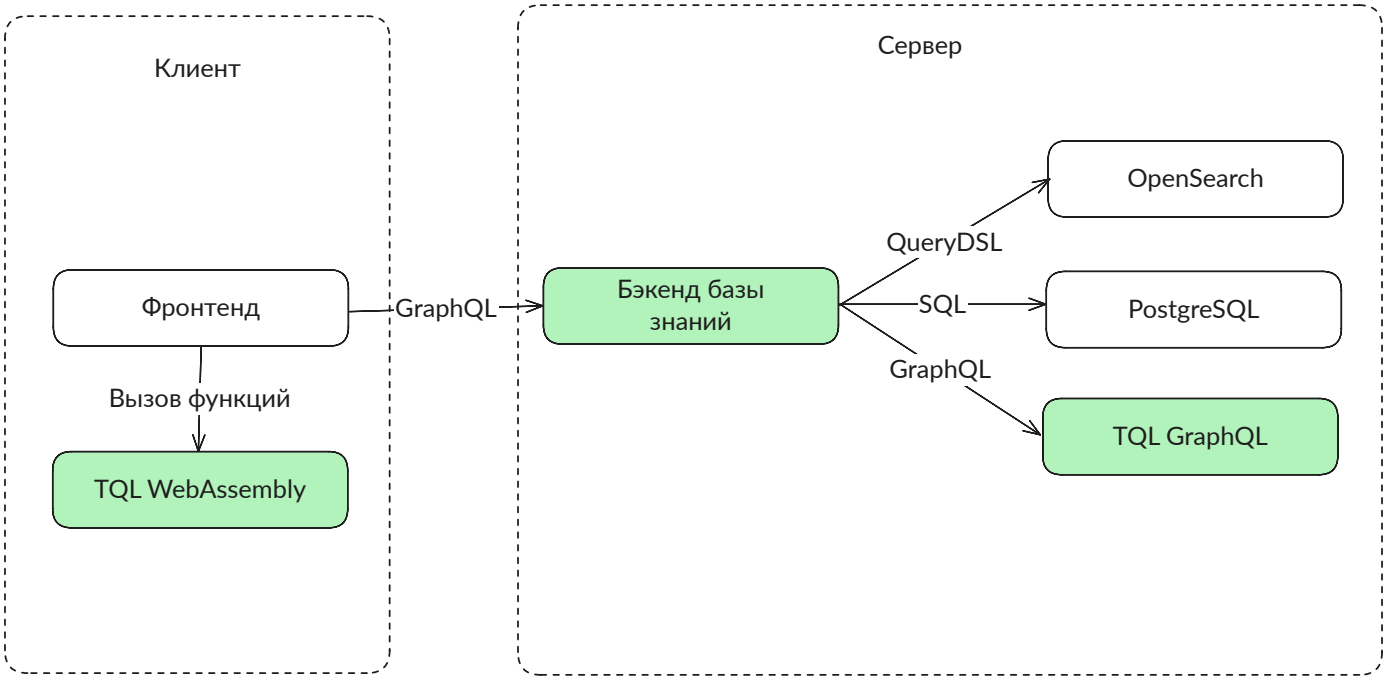
Что будет найдено по запросу?

```
Концепт(город=Переславль),Концепт(город!=Переславль)
```

Инструментарий для TQL

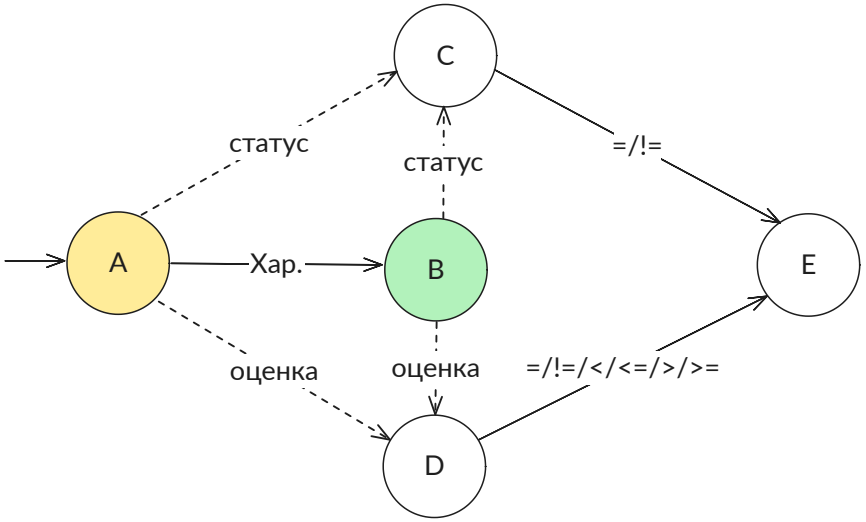
- Автодополнение
  - Ключевые слова языка
  - Данные предметной области
- Валидация запросов
  - Синтаксис языка
  - Семантика языка
  - Семантика предметной области
- Перевод запросов

Архитектура поиска в Talisman



Автодополнение

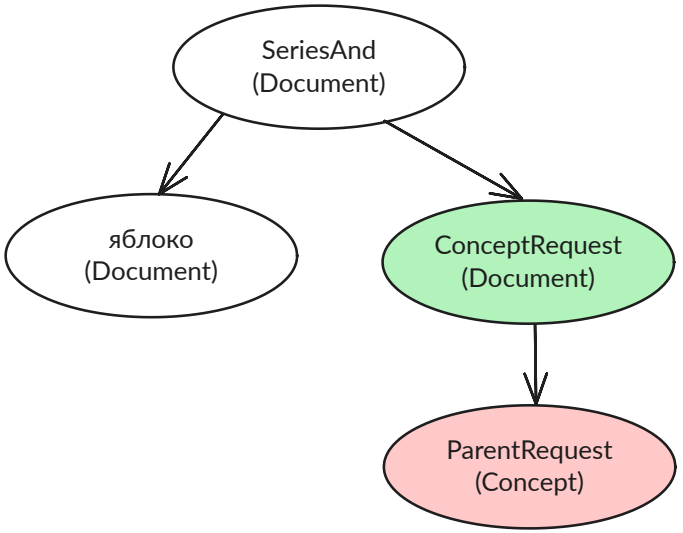
Запрос: Хар.ста



Дополнения: Хар.статус=, Хар.статус!=

Валидация

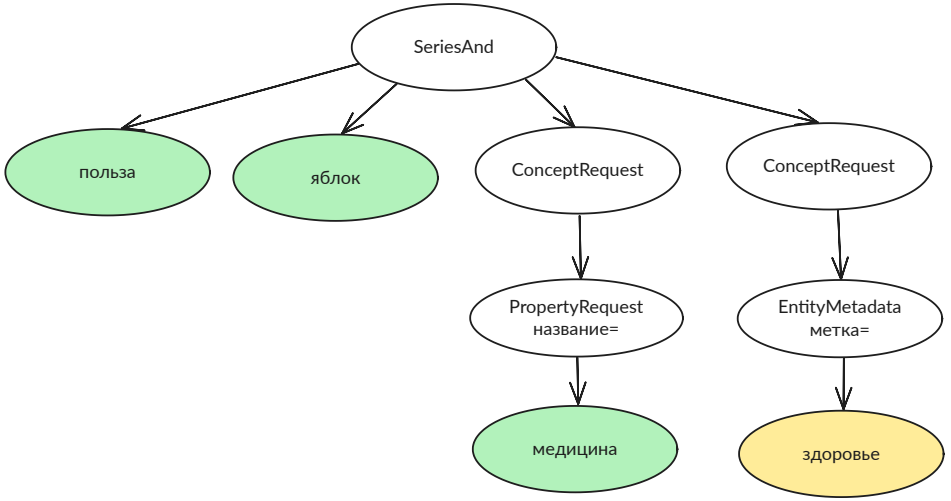
Запрос: яблоко Концепт (Родительский ())



Ошибка: селектор «Родительский» недоступен в контексте «Концепт»

Перевод текстовых элементов

Запрос: польза яблок Концепт(название=медицина) Концепт(метка=здоровье)



Перевод:
The benefits of apples Концепт(название=medicine) Концепт(метка=здоровье)

Итого

- Выпускникам ИУ9 стоит рассмотреть программу «Системное программирование» ВШЭ
- При решении разрабатывать свой DSL следует учитывать не только сложность разработки языка, но и инструментария
- Невозможно угадать, где пригодятся полученные навыки и знания