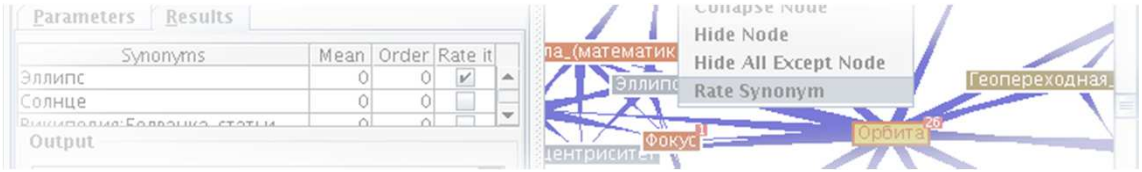


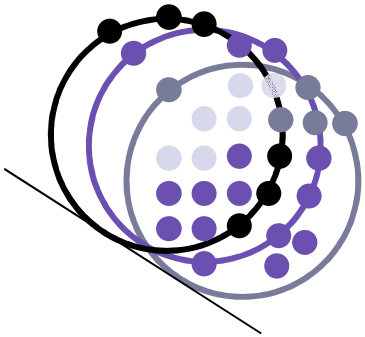
Multilingual Ontology Matching based on Wiktionary Data Accessible via SPARQL Endpoint

Санкт-Петербургский институт информатики и автоматизации РАН

Фейю Лин feiyu.lin jth.hj.se

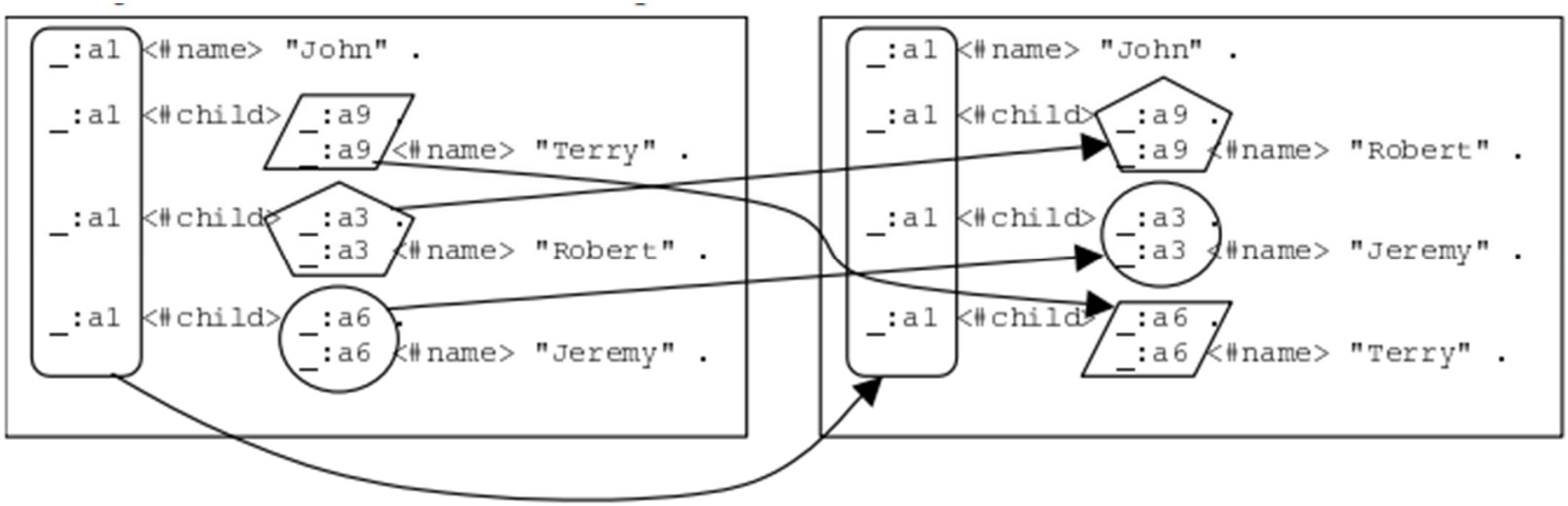
Крижановский Андрей andrew.krizhanovsky gmail.com

Содержание



- Ontology matching
- Викисловарь
- SPARQL
- Отображение онтологий на разных языках

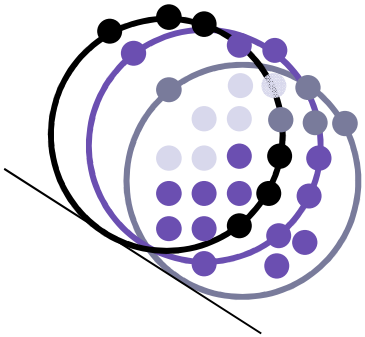
Multilingual Ontology matching



RDF / XML файл 1

RDF / XML файл 2

Постановка задачи



- Выполнить отображение онтологий на разных языках (англ., фр.)
- Перевести с помощью:
  - Google Translate API
  - **Викисловарь** (машинно-читаемый словарь)

Сравнить качество перевода

Викисловари

Викисловарь –

<u>многофункциональный</u>

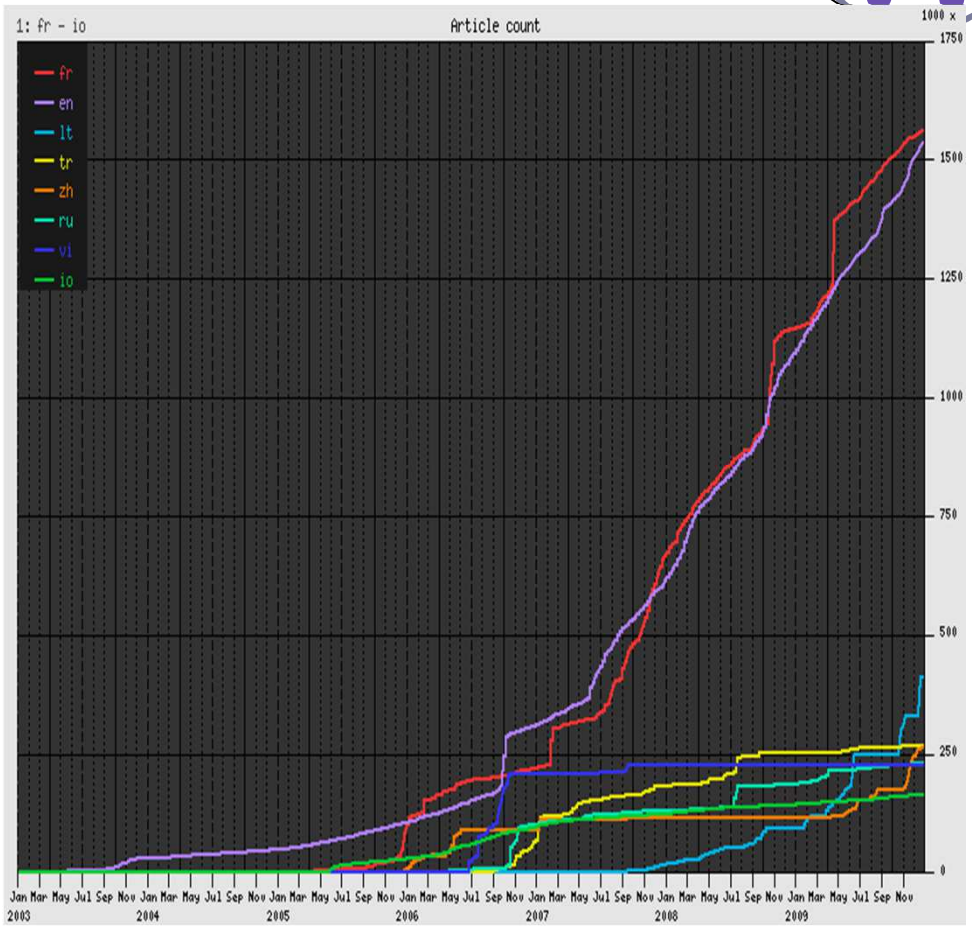
многоязычный

словарь и тезаурус

Грамматический
Толковый
Этимологический
Переводной

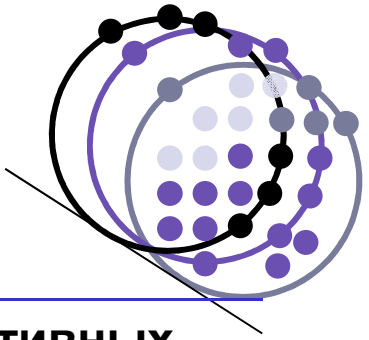
Развитие Викисловарей

+ Первым появился English Wiktionary в декабре 2002 г.
+ Проект Русский Викисловарь запущен в мае 2004 г.



Восемь самых больших Викисловарей (2003-2010)

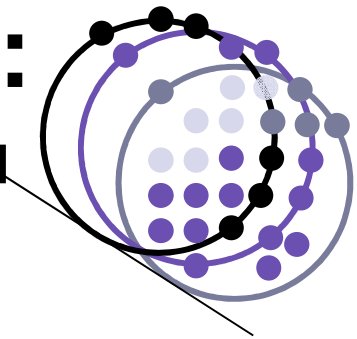
10 крупнейших* (из 170) Викисловарей



N	Языковая версия	Словарных статей	Администра-торов	Активных редакторов
1	**Английский Викисловарь**	**2 639 367**	**90**	**980**
2	Французский	2 081 306	23	314
3	Китайский	1 197 238	8	44
4	Малагасийский	739 456	2	8
5	Литовский	562 276	4	22
6	***Русский***	***300 332***	***7***	***174***
7	Турецкий	278 631	6	55
8	Польский	250 385	26	85
9	Тамильский	234 568	12	47
10	Корейский	229 410	1	20

* По данным на октябрь 2011

Английский Викисловарь:
Число словарных статей по языкам (Многоязычность)

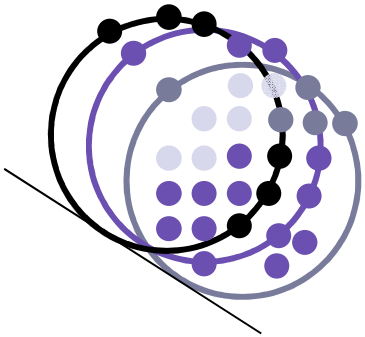


Раздел Английского Викисловаря	Число словарных статей
Латинский	377 791
Итальянский	318 707
Английский	*304 483*
Французский	132 423
Финский	97 741
и так далее…	

- Словарные статьи о словах 433 языков.
- Переводы английских слов – на 235 языков.

Дамп словаря от 18 июня 2011 г.

Машинно-читаемый словарь на основе Английского Викисловаря: Раздел переводов



	Английский Викисловарь	Machine-readable (MRD)
Словарных статей	2 651 524	2 149 177
Число переводов с Английского на другие языки	?	756 168
Языков (> 1 000 переводов)	?	68
Формат данных	Разметка Wiki / XML	Реляционная БД
Противоречия и ошибки ввода данных	Есть	Нет (проверка структуры статьи парсером)
Программный интерфейс (API) для перевода слова	Нет	Есть

encyclopedia
Wiktionary
['wɪkʃənrɪ] *n.*,
a wiki-based Open Content dictionary

search

Go Search

navigation

- Main Page
- Community portal
- Preferences
- Requested entries
- Recent changes
- Random entry (by language)
- Help
- Donations
- Contact us

French

Noun

fleur *f* (*plural* **fleurs**)

1. (*botany*) Flower; bloom; blossom; collectively, the reproductive organs and the envelope which surrounds them in angiosperms (also called "flowering plants").

   *Je suis allé cueillir une **fleur** dans les champs.*

   I went to pick a flower in the fields.

Synonyms

- (*flowering plant*): angiosperme

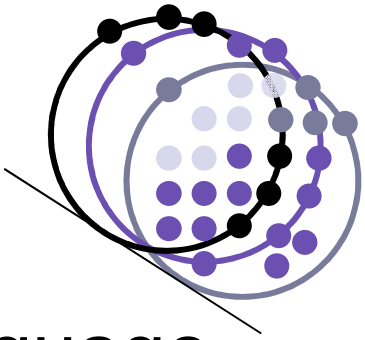
Hyponyms

- (*flower, bloom, blossom*): bractée, carpelle, étamine, fleuron, pédoncule, pétale, pistil, sépale, tépale

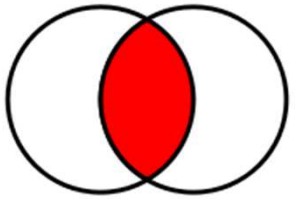
An example of epibolium flowers (*fleurs d'épilobes*)

Слово	Язык (код)	Часть речи	Число сем. отн.	Число типов сем. отн.	Число значений
fleur	French (fr)	Noun	25	4	6

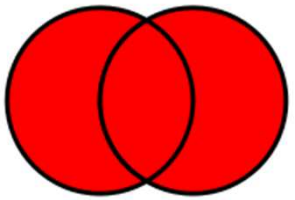
SPARQL



- **S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage
- RDF язык запросов:
  - Тройка, триплет (?X Отношение ?Y)
  - Конъюнкция (A **;** B.)
    - SELECT ?x WHERE { ?A Relation1 ?x **;**
      ?x Relation2 ?B **.** }

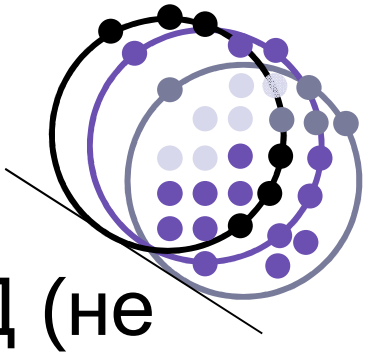


  - Дизъюнкция (A**.** B**.**)
    - SELECT ?x WHERE { ?A Relation1 ?x **.**
      ?x Relation2 ?B **.** }



  - Дополнительные шаблоны для уточнения поиска.

Платформа D2RQ



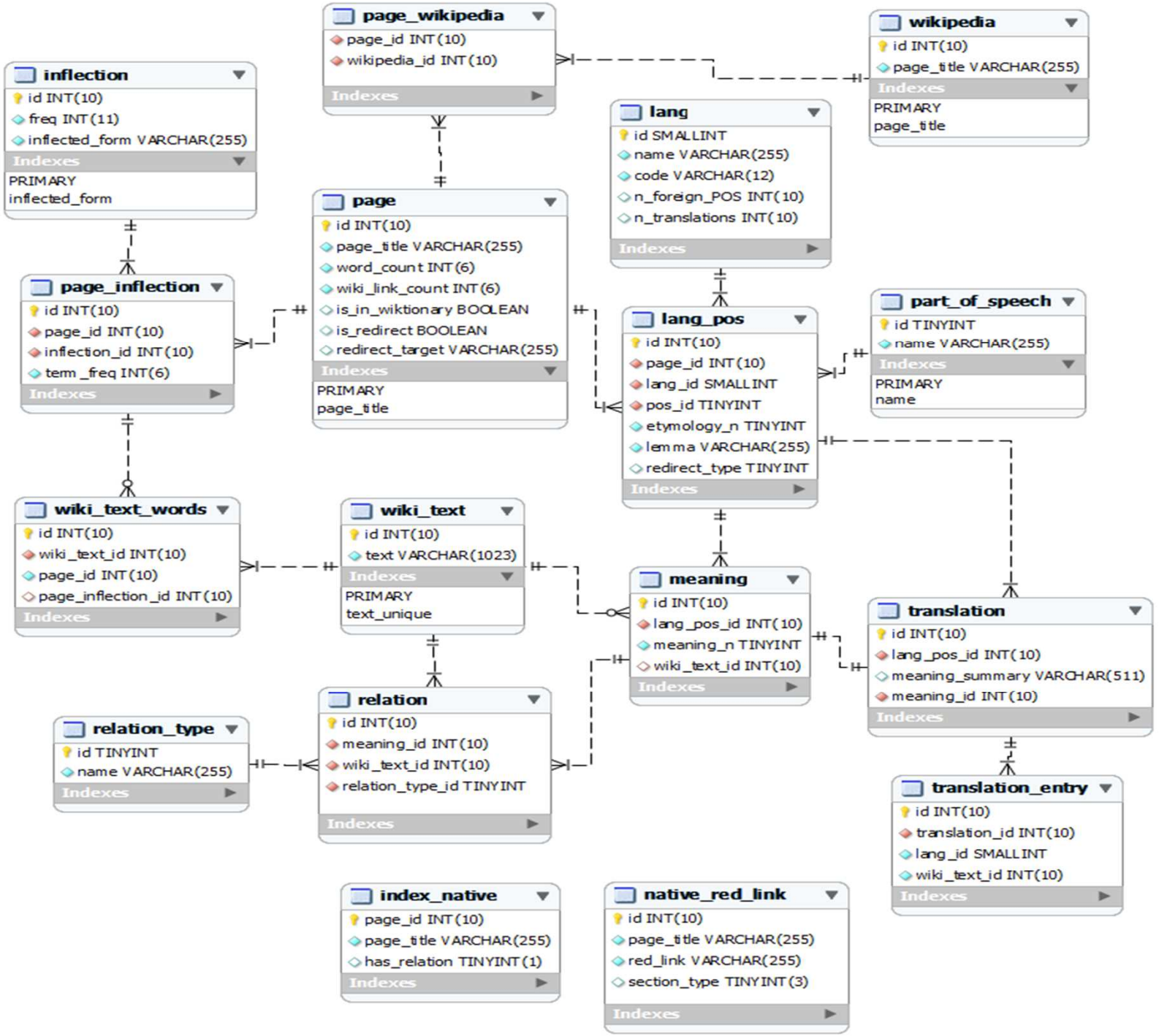
- D2RQ представляет реляционную БД (не RDF) как виртуальный RDF граф
  1. Создать файл отображения (MySQL - RDF)
  2. Запустить MySQL и сервер-D2RQ
  3. Запустить запросы SPARQL
- Инструкции: D2RQ и данные Викисловаря

  http://code.google.com/p/wikokit/wiki/d2rqMappingSPARQL

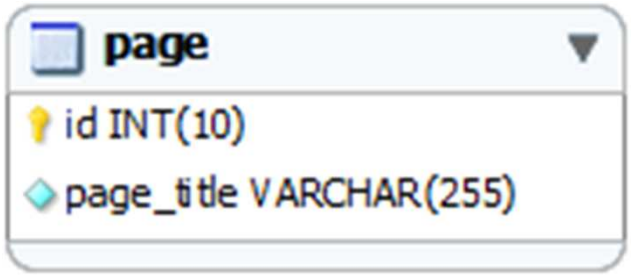
Wik-tion-ary

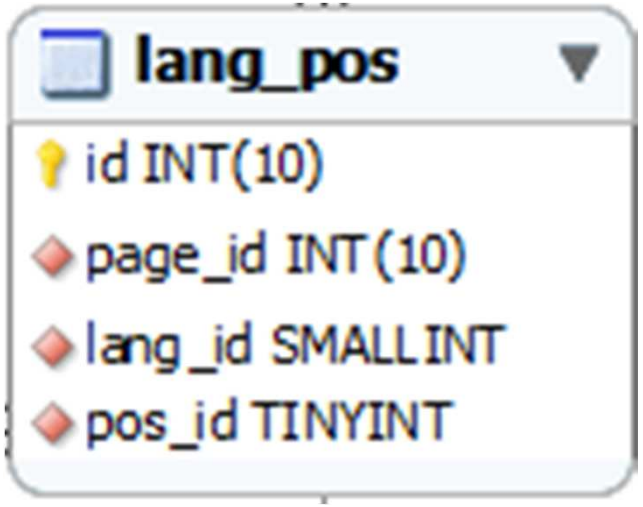
MRD

data-base

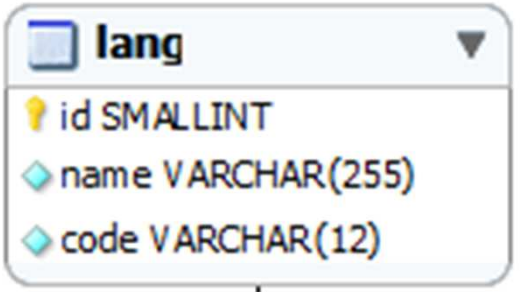


Получить толкование из Викисловаря по слову и языку (SPARQL запрос)

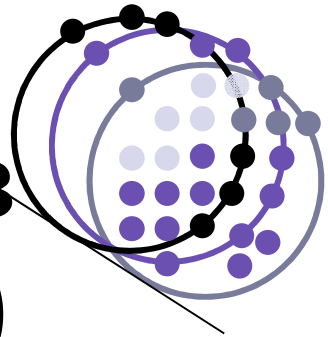




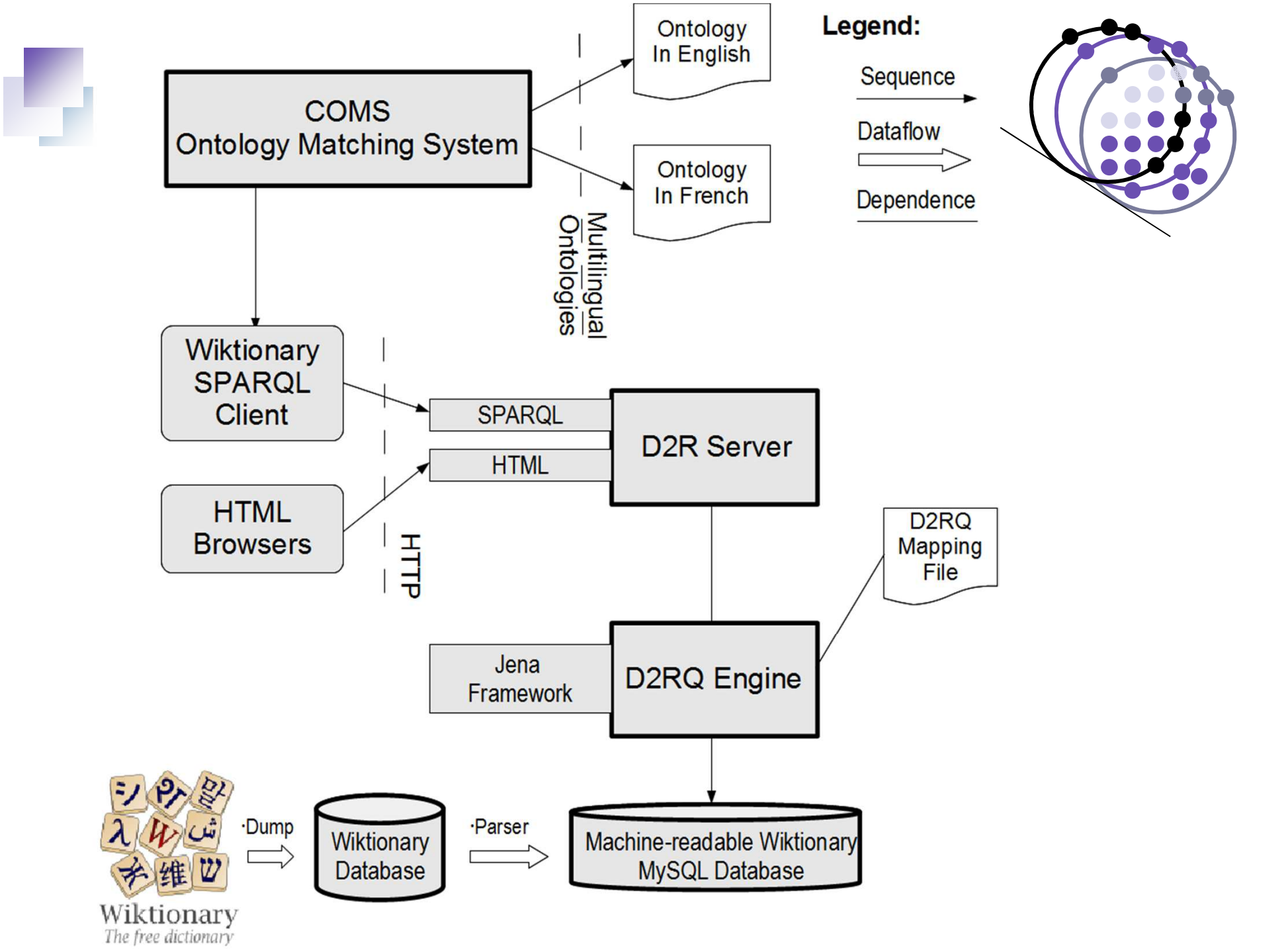

```
SELECT ?langId ?pageId ?langPosId ?meaningId
?wikiTextIdDef ?definition
WHERE {
    ?lang wikpa:lang_code "en";
          wikpa:lang_id ?langId.

    ?page wikpa:page_page_title "dog";
          wikpa:page_id ?pageId.

    ?lang_pos wikpa:lang_pos_page_id ?pageId;
              wikpa:lang_pos_lang_id ?langId;
              wikpa:lang_pos_id ?langPosId.

    ?meaning wikpa:meaning_id ?meaningId;
             wikpa:meaning_lang_pos_id ?langPosId;
             wikpa:meaning_wiki_text_id ?wikiTextIdDef.

    ?wiki_text wikpa:wiki_text_id ?wikiTextIdDef;
               wikpa:wiki_text_text ?definition.
}
```



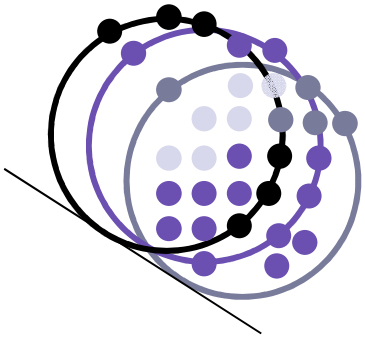
Ответ-SPARQL (список толкований слова “dog” из Английского Викисловаря)



langId	pageId	langPosId	meaningId	wikiTextIdDef	definition
262	362	8	26	353	"An animal, member of the genus "Canis" (probably descended from the common wolf) that has been domesticated for thousands of years; occurs in many breeds. Scientific name: "Canis lupus familiaris"."
262	362	8	27	508	"A male dog, wolf or fox, as opposed to a bitch (a female dog, wolf or fox.)"
262	362	8	28	524	"{{derogatory}} A dull, unattractive girl or woman."
262	362	8	29	528	"{{slang}} A man."


COMS
Ontology Matching System
Ontology
In English
Ontology
In French
Multilingual
Ontologies
Legend:
Sequence
Dataflow
Dependence
Wiktionary
SPARQL
Client
HTML
Browsers
SPARQL
HTML
D2R Server
HTTP
D2RQ
Mapping
File
Jena
Framework
D2RQ Engine
·Dump
Wiktionary
Database
·Parser
Machine-readable Wiktionary
MySQL Database
Wiktionary
The free dictionary

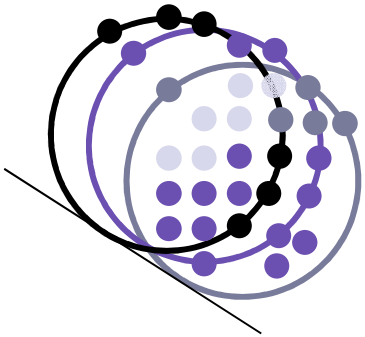
Эксперимент



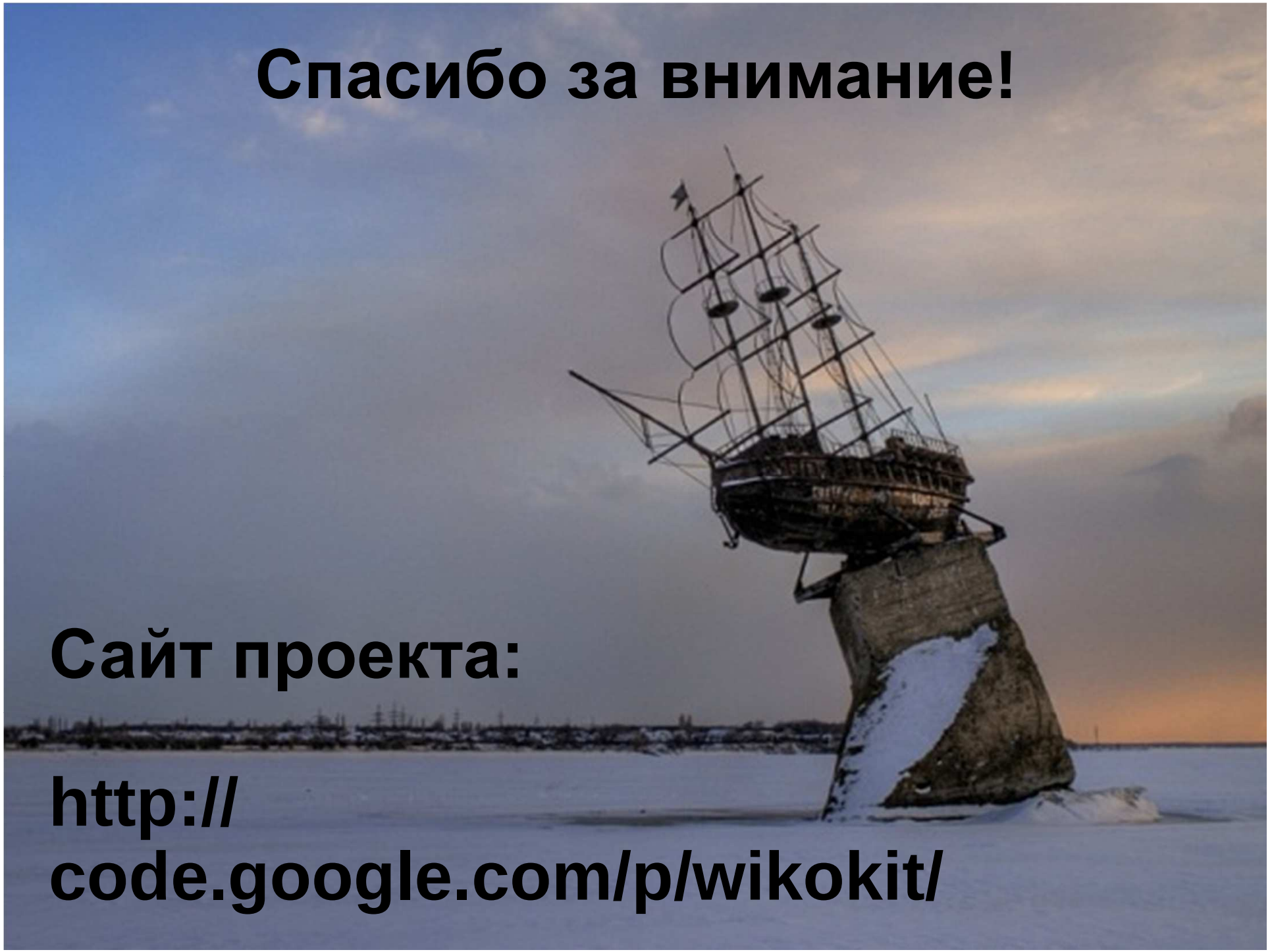
- Тестовые данные (OAEI)
  - Онтология на англ. и на фр.
    - На фр. языке: 85 классов, 97 атрибутов = 182
- Идеальное отображение - 97 элементов

	Правильных переводов	Число элементов в отображении онтологий	Precision	Recall
MRD Wiktionary	44	54	0.98	0.55
Google	60	61	0.98	0.62

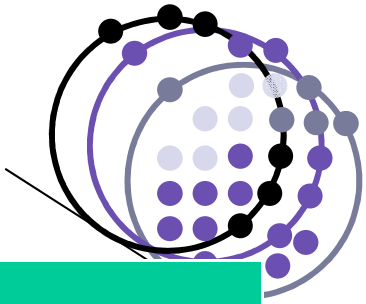
Результаты



- SPARQL запросы к Викисловарю
  - Список толкования по слову и языку
  - Список синонимов
  - Перевод слова
    (с английского на один из языков)
    - http://code.google.com/p/wikokit/wiki/d2rqMappingSPARQL
- Пример приложения на Java
  - создаёт SPARQL запросы и получает данные от D2RQ сервера


Спасибо за внимание!
Сайт проекта:
http://
code.google.com/p/wikokit/

Машинно-читаемый Викисловарь (MRD): Синонимы



	Викисловарь*	MRD
Словарных статей	2 651 524	2 149 177
Число языков	288	
Число языков (>10 000 словарных статей)	26	
Число языков с парадигм. отношен.**	?	235
Языков (> 1 000 отн.)	?	26
Число сем. отношений	?	220 211
Формат данных	Разметка Wiki / XML	Реляционная БД
Противоречия и ошибки ввода данных	Есть	Нет

*Английский Викисловарь, дамп от 18 июня 2011 г.

** Парадигматические (семантические) отношения – синонимы, антонимы, гиперонимы…

Схема отображения онтологий