



МИНИСТЕРСТВО НАУКИ
И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ



СПб
ФИЦ
РАН

Санкт-Петербургский
Федеральный исследовательский центр
Российской академии наук



Модели и методы онтолого-ориентированной интеграции информации из разнородных источников с учетом пространственно-временных отношений для задач геоанализа

Тесля Николай Николаевич
к.т.н., с.н.с. лаб. ИСА, СПИИРАН
СПб ФИЦ РАН



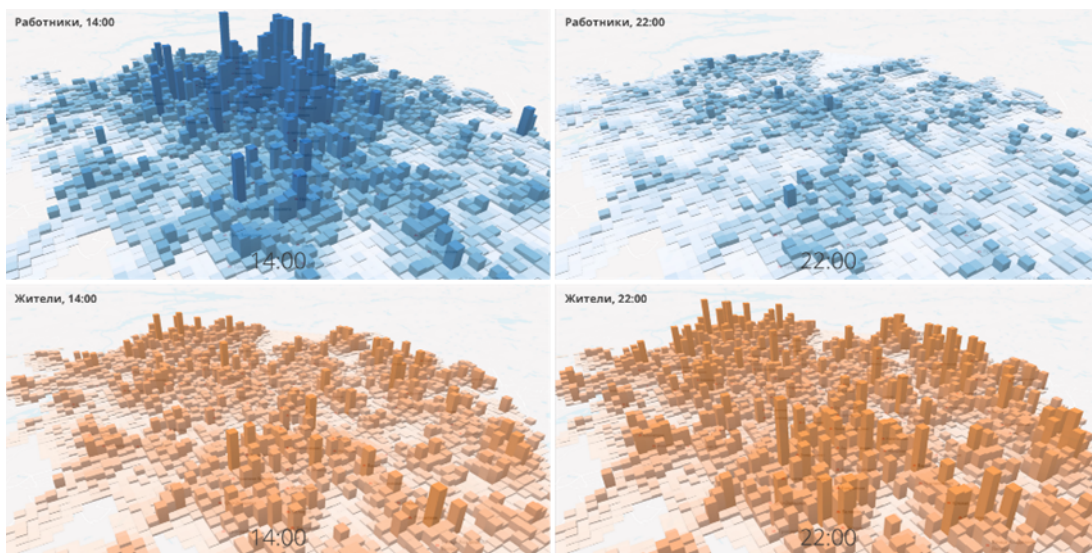
Структура

- Геоанализ
- Источники данных и знаний, имеющие привязку к пространству и времени
- Моделирование пространственно-временных отношений в онтологиях
- Метод формирования общей онтологии из нескольких источников данных и знаний
- Модель сервиса для поиска информации, имеющей привязку к пространству и времени
- Методы поиска информации в источниках знаний, имеющих привязку к пространству и времени
 - GeoSPARQL и федеративный SAPRQL
- Пример геоанализа для карточек ДТП



Геоанализ

- Анализ данных, имеющих привязку к пространству и времени
- Источники данных, как правило, геоинформационные сервисы, как государственные, так и коммерческие.
- Цель анализа – определить закономерности распределения данных в пространстве и, по возможности, отследить их динамику во времени.
- Пример – Анализ маятниковых миграционных потоков в городе.

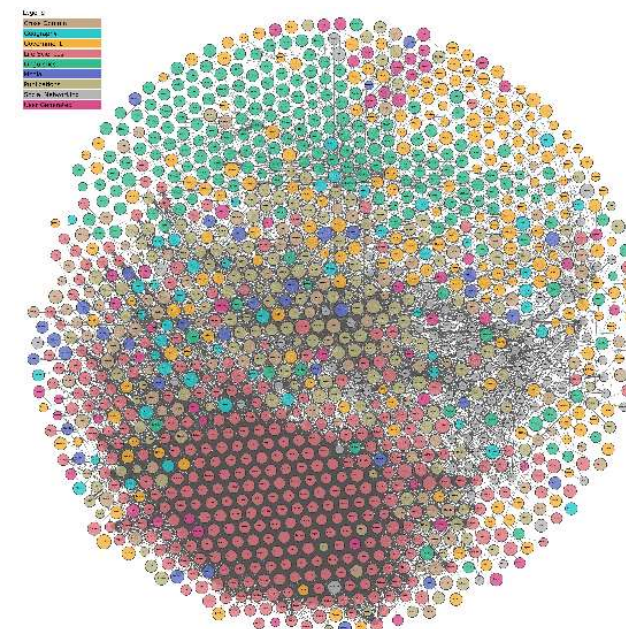


Источник: Дизайн города, основанный на данных
<https://habr.com/ru/post/398773/>

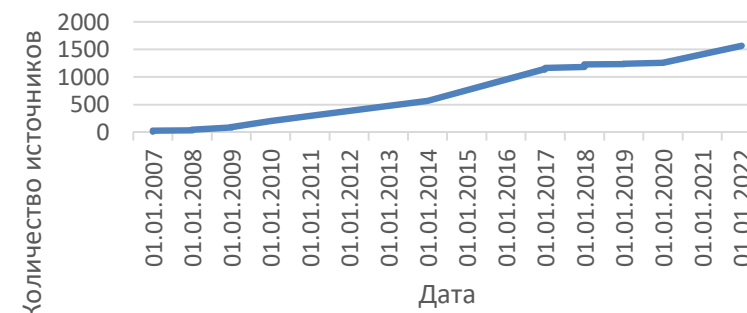


Мотивация

- Поиск информации, совмещающий пространственные и временные характеристики для геоанализа, должен осуществляться не только по точному совпадению координат или временных отметок, но и с учетом семантики отношений в пространственной и темпоральной логики (вывод на отношениях типа «рядом», «около», «до», «после» и т.д.)
- Существующие системы работают с **точными** соответствиями объектов системы и запроса.
- Семантическая модель данных (онтологи) объединяет пространственные и временные концепты в единой схеме данных, но позволяет учесть как пространственные, так и временные отношения при поиске.
- Связанные (linked) данные содержат связи между отдельными наборами данных, создавая единую связную онтологию.
- Формируются репозитории связанных (linked) данных.
 - <https://lod-cloud.net/> - репозитарий, содержащий 1568 наборов связанных данных.
 - 2007 – 2 млрд RDF-троек и 2 млн ссылок RDF.
 - 2011 – 31 млрд RDF-троек и 504 млн ссылок RDF.
- Возможно преобразование реляционных моделей в онтологические посредством медиатора.



Динамика добавления источников связанных данных в репозитарий lod-cloud.net





Контекстно-ориентированная интеграция информации для геоанализа

- Объем данных в открытом доступе в формате онтологии (RDF и OWL) с привязкой к пространству и времени постоянно растет.
- **Контекст** в данном случае понимается как онтолого-ориентированная модель представления данных и знаний, содержащая **актуальные** свойства концептов.
- Контекст может быть составлен в результате интеграции нескольких источников данных и информации, что предоставляет возможность контекстно-ориентированного управления информацией.
- Онтология верхнего уровня при интеграции информацией объединяет в себе концепты из различных предметных областей, относящиеся к одной из пяти фундаментальных категорий: действие, местоположение, отношение, время, индивидуальность (тип).
- Источники информации и сама информация при расширении контекста в задаче геоанализа могут подбираться:
 - в соответствии с **координатами** рассматриваемого концепта;
 - в соответствии с **временными** отметками;
 - в соответствии с пространственно-временными **отношениями** концептов;
 - в соответствии с **типом** информации.



Фундаментальные категории контекстной информации



Основные области применения геоанализа

- Отслеживание изменения местности (с/х, урбанистика, геодезия, геология, ДЗЗ)
- Отслеживание истории погоды и климатических изменений в привязке к координатам (прогнозирование погоды)
- Отслеживание изменений и характеристик городской среды («умный город»)
- Анализ изменения качества покрытия радиосвязи и изменения пропускной способности радиоканалов (мобильная, радио-, спутниковая связь)
- Анализ популяций и миграций птиц и животных по радиометкам.
- Расширение контекста объекта или события информацией из дополнительных источников.
- Верификация данных из различных источников.

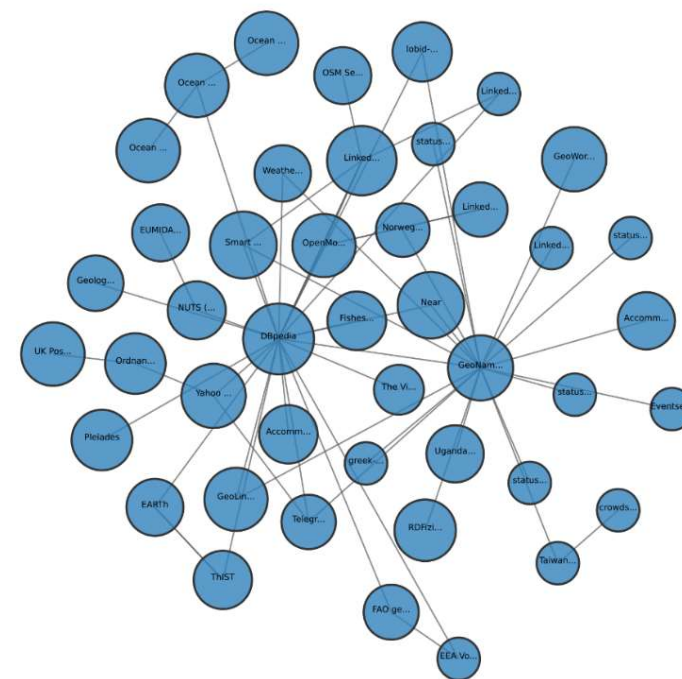


Примеры онтологических источников знаний, имеющих привязку к пространству и времени

- **DBpedia** – данные из Википедии. Содержит около 33.8 млн концептов (23.8 млн – локализованные версии существующих концептов) описанные 21 млрд триплетов на 125 языках. (2020)
- **GeoNames** – 11 млн топонимов по всему миру. (2022)
- **Wikidata** – совместно созданный связанный набор данных, который действует как центральное хранилище для структурированных данных проектов Фонда Викимедиа. Содержит около 100 млн объектов. (2019)
- **KnowWhereGraph** граф знаний с пространственной привязкой. Содержит около 12 млрд триплетов.
- **LinkedGeoData** содержит около 3 млрд узлов и 300 млн путей. Основан на данных OpenStreetMap, и делает ее доступной в виде базы знаний RDF в соответствии с принципами Linked Data (примерно 20 млрд троек). Он связывает эти данные с другими базами знаний в рамках инициативы Linking Open Data. (с 2012 г.)

Legend

Geography



The Geography Linked Open Data Cloud from lod-cloud.net





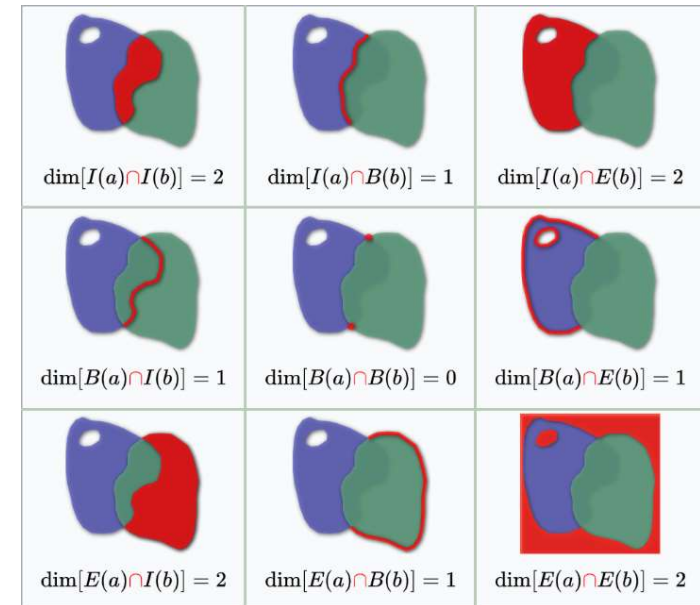
Примеры реляционных и иных источников данных, имеющих привязку к пространству и времени

- Open Street Map
- Порталы открытых данных (ОД). Например, портал ОД России, СПб, Москвы и т.д.
- Географические БД:
 - SRTM (Shuttle Radar Topography Mission) — цифровая модель рельефа (<https://ers.cr.usgs.gov>)
 - Свободные данные по границам субъектов РФ (Росреестр + OSM + GISLAB <https://gis-lab.info/qa/rusbounds-rosreestr.html>)
- Архивы метеоданных:
 - NOAA National Centers for Environmental Information (NCEI)
 - OpenWeatherMap - метеорологический архив по координатам с интервалом 1 час
 - «rp5.ru Расписание погоды» - метеорологический архив по городам с интервалом в 3 часа.
- Исторические карты (требуют дополнительной оцифровки)



Моделирование пространственных отношений в онтологии

- До сих пор (по состоянию на 10 марта 2023) не существует единой согласованной онтологии для описания пространственных концептов (<https://w3c.github.io/sdw/qb4st>).
- Сложность заключается в составлении словаря концептов, их свойств и отношений между ними, а также, формализации операторов (например, для 2D пространства ~ 512 топологических отношений, обобщенные в 9 классов).
- Отдельной задачей является формализация механизма логического вывода над пространственными знаниями.
- Наиболее часто используемая схема для представления пространственных данных: ISO 19107:2019 Geographic information — Spatial schema.
- В онтологии частично проработаны базовые концепты и отношения, например, в GeoSPARQL представлены базовые концепты пространства.
- В ГИС для описания отношений используется топологическая модель DE-9IM (**D**imensionally **E**xtended **9-Intersection Model**)

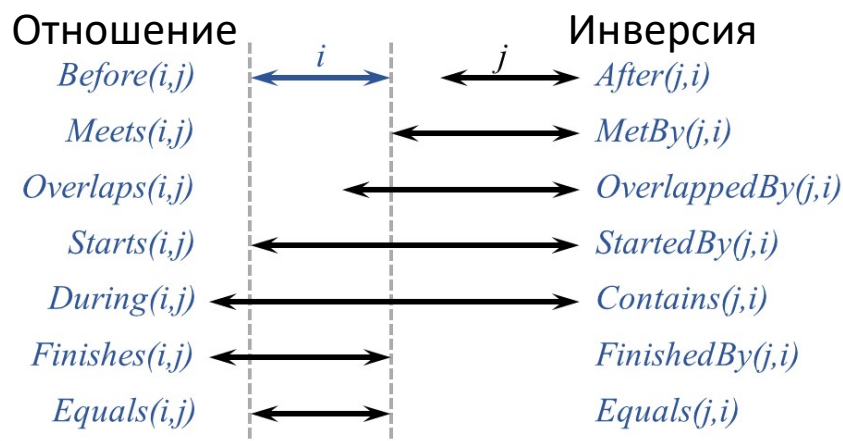


Предикаты для двух областей:
Равенство, непересечение,
касание, содержание,
покрытие, пересечение,
вложенность, обратное
покрытие



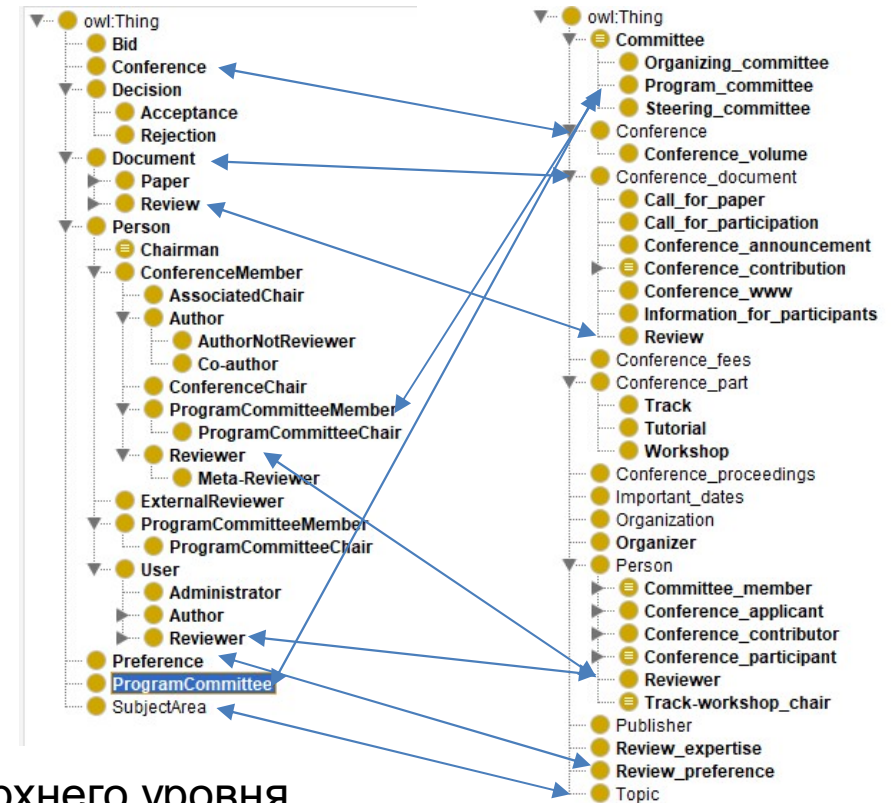
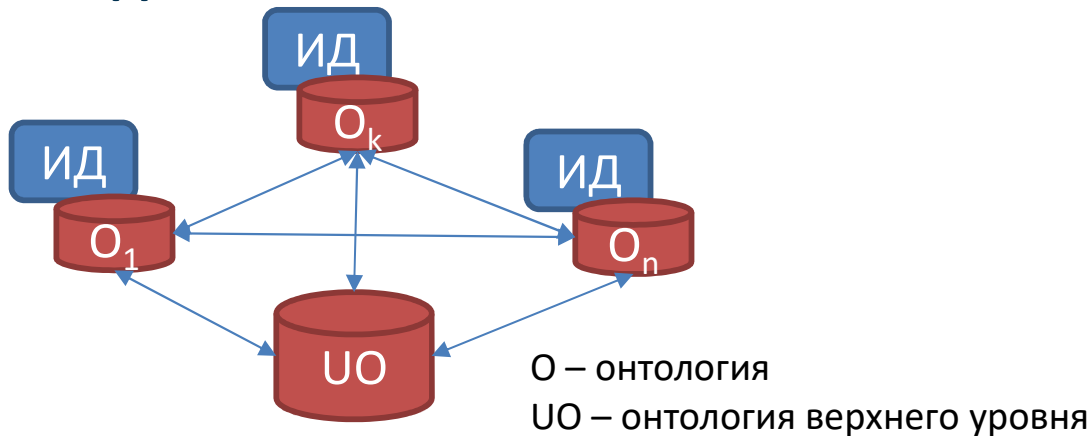
Моделирование временных отношений в онтологии

- В отличие от пространственных отношений существует рекомендация от W3C для описания временных отношений – OWL-Time-ontology (<https://www.w3.org/TR/owl-time>, 2020)
- OWL-Time-ontology основана на алгебре бинарных отношений на интервалах (пересечение, перекрытие, вложенность и т.д.) (J.F. Allen, Towards a general theory of action and time. Artificial Intelligence. Vol. 23, iss. 2, 1984, pp. 123-154) для представления различных типов событий, процессов, действий и свойств на английском языке.
- Для русского языка данные отношения подробно описаны в Кандрашина Е.Ю., Литвинцева Л.В., Поспелов Д.А. Представление знаний о времени и пространстве в интеллектуальных системах / Под ред. Д.А. Поспелова - М.: Наука, 1989. - 328 с.
- Для логического вывода используется темпоральная логика, представленная с использованием логики предикатов.





Метод формирования общей онтологии из нескольких источников данных и знаний



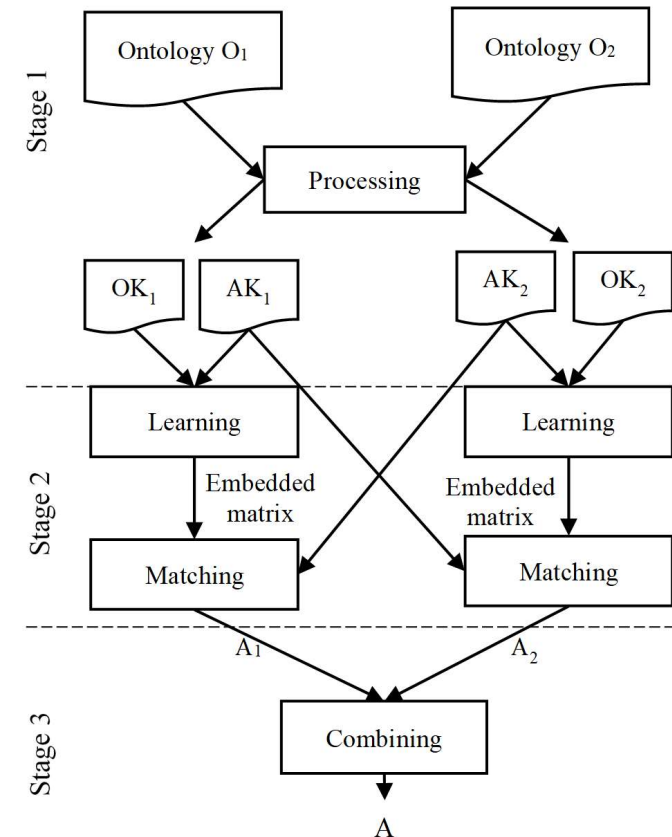
- Сопоставление онтологий
 - Создание общей онтологии
 - Связывание частных онтологий с онтологией верхнего уровня
- Подходы:
 - Основанные на анализе элементов (контент)
 - Основанные на анализе структуры (контекст)
- Предложен метод на основе нейронных сетей, совмещающий оба подхода.



Нейросетевой метод сопоставления онтологий

- Объединение элементного и структурного анализа.
- Входной вектор – название концепта, связанные концепты, типы связей (абстрактный контекст), характеристики концепта (оперативный контекст)
- 2 нейронных сети:
 - Анализ семантического соответствия названий концептов: Word2Vec NN (библиотека SpaCy)
 - Анализ структуры и характеристики (Convolutional NN)
- Результат работы – принадлежность рассматриваемого концепта онтологии (классификация) и оценка семантической меры сходства.

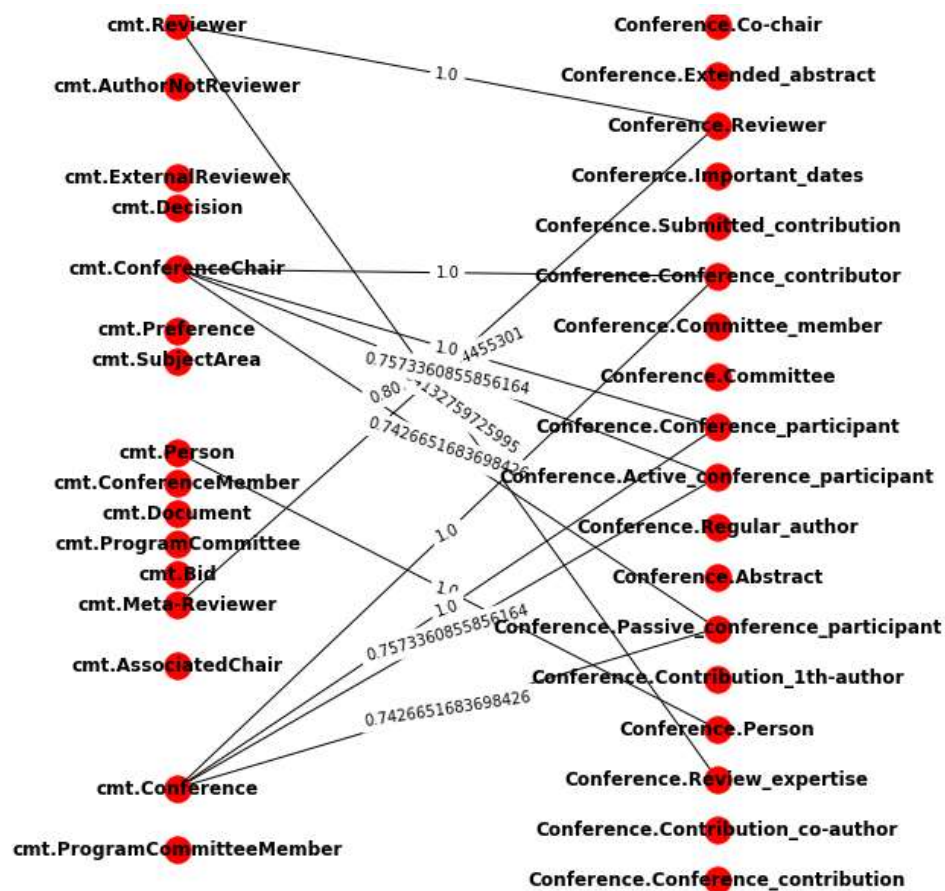
Метрика	Предложенный подход	ALOD2Vec	StringEquiv	AML
Precision	0.79	0.71	0.8	0.84
Recall	0.65	0.5	0.43	0.66
F1-measure	0.71	0.59	0.56	0.74



Данные для сравнения:
OAEU 2019 – conference -
ra1-M3 (16 онтологий, в
среднем 54 класса)



Результат сопоставления





Преобразование реляционных источников данных в онтологию

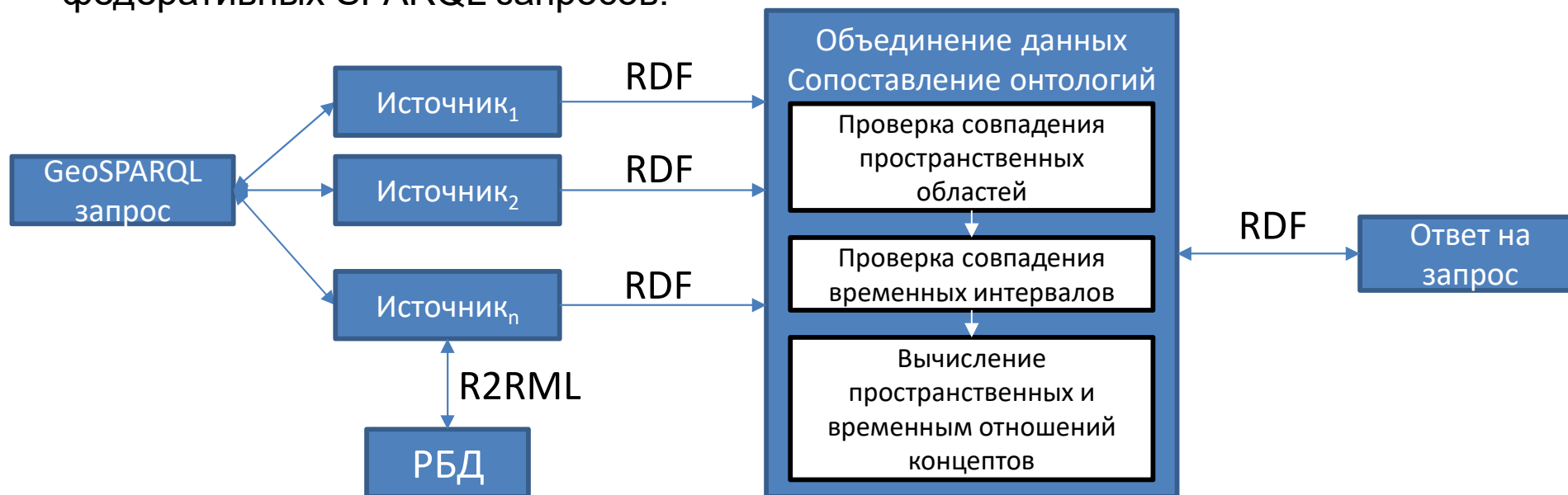
- R2RML* – язык для преобразования реляционной БД в набор данных RDF.
- Платформы:
 - -ontop- создает виртуальные RDF-графы из реляционных данных и поддержкой запросов на языке SPARQL.
 - *OpenLink Virtuoso*
- Порядок преобразования:
 - БД приводится в 3 нормальную форму.
 - Связи таблиц преобразуются в отношения.
 - Таблицы преобразуются в концепты, поля таблиц – в свойства концептов.
 - Формируются RDF-тройки, соответствующие строкам таблиц и отношениям между таблицами.
- К полученному преобразованию может применяться SPARQL для поиска данных.

*<https://www.w3.org/TR/r2rml/>



Поиск данных из разнородных источников с привязкой к пространству и времени

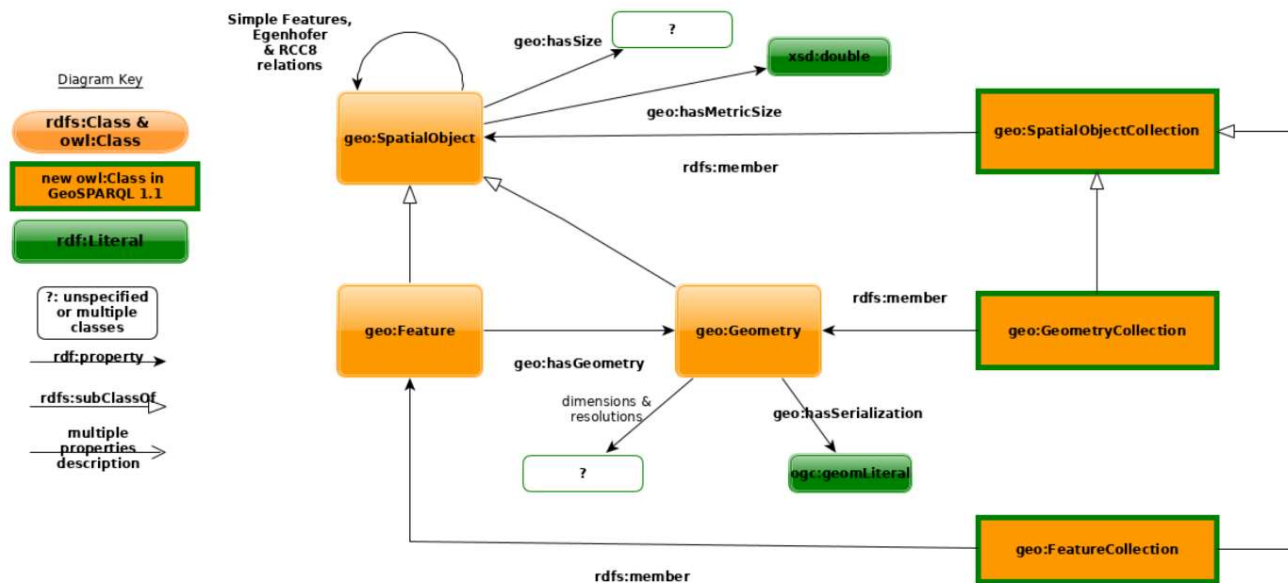
- R2RML не поддерживает автоматическую трансляцию пространственных объектов и отношений из реляционной модели в онтологию.
- На текущий момент транслируются только базовые геометрии, при условии их явного указания в коде трансляции вручную.
- GeoSPARQL – Язык представления и запросов к пространственным связанным данным.
- Для использования нескольких источников данных в SPARQL существует стандарт федеративных SPARQL запросов.





GeoSPARQL

- Расширение языка SPARQL
- Содержит базовую онтологию представления топологии.
- Содержит базовые топологические отношения (DE-9IM, RCC8*).
- Последняя версия языка – 1.1 (2021 год**)

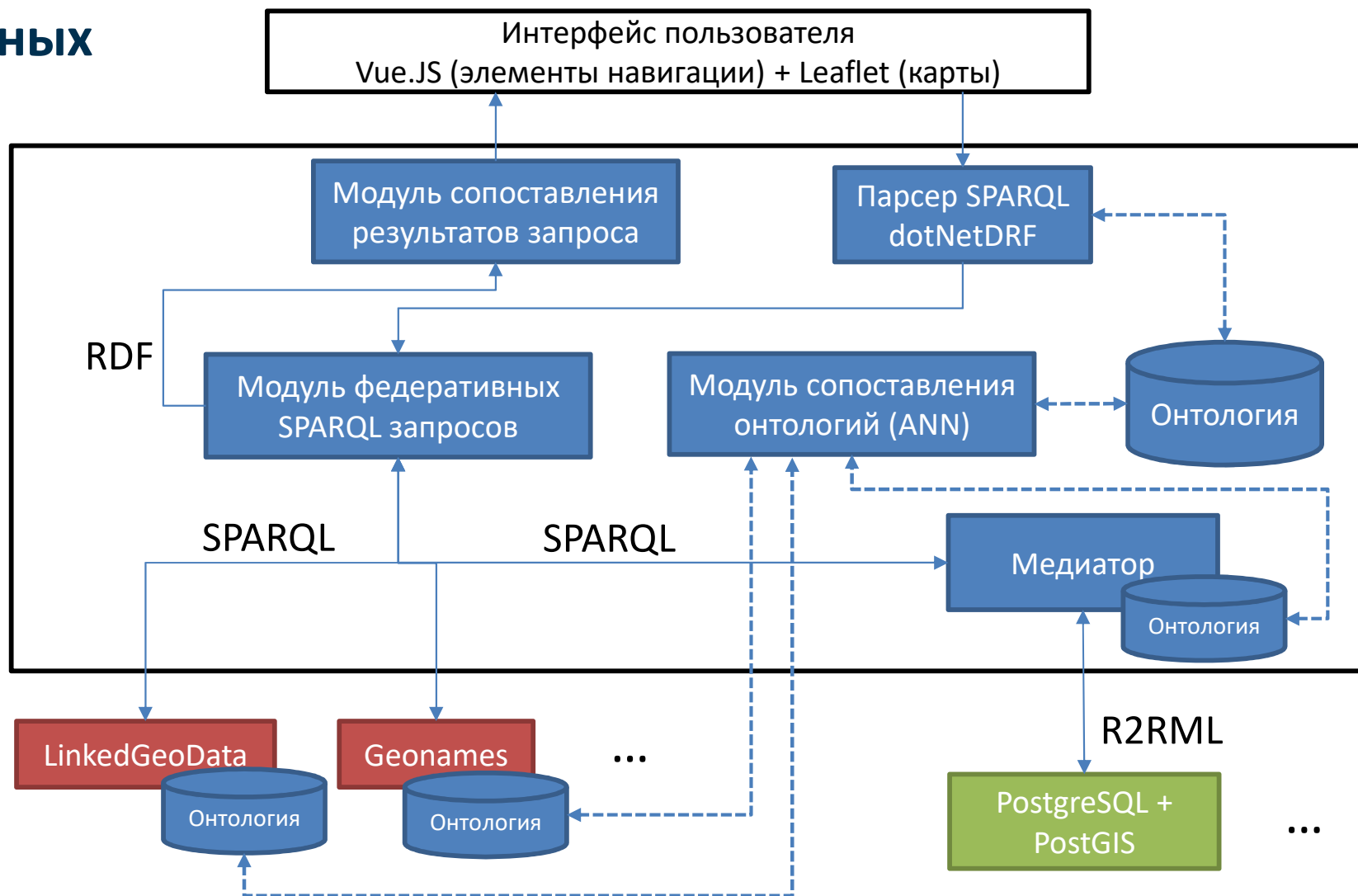


*Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. *KR*, 92, 165-176.

** Car, N.J.; Homburg, T. GeoSPARQL 1.1: Motivations, Details and Applications of the Decadal Update to the Most Important Geospatial LOD Standard. *ISPRS Int. J. Geo-Inf.* **2022**, 11, 117. <https://doi.org/10.3390/ijgi11020117>

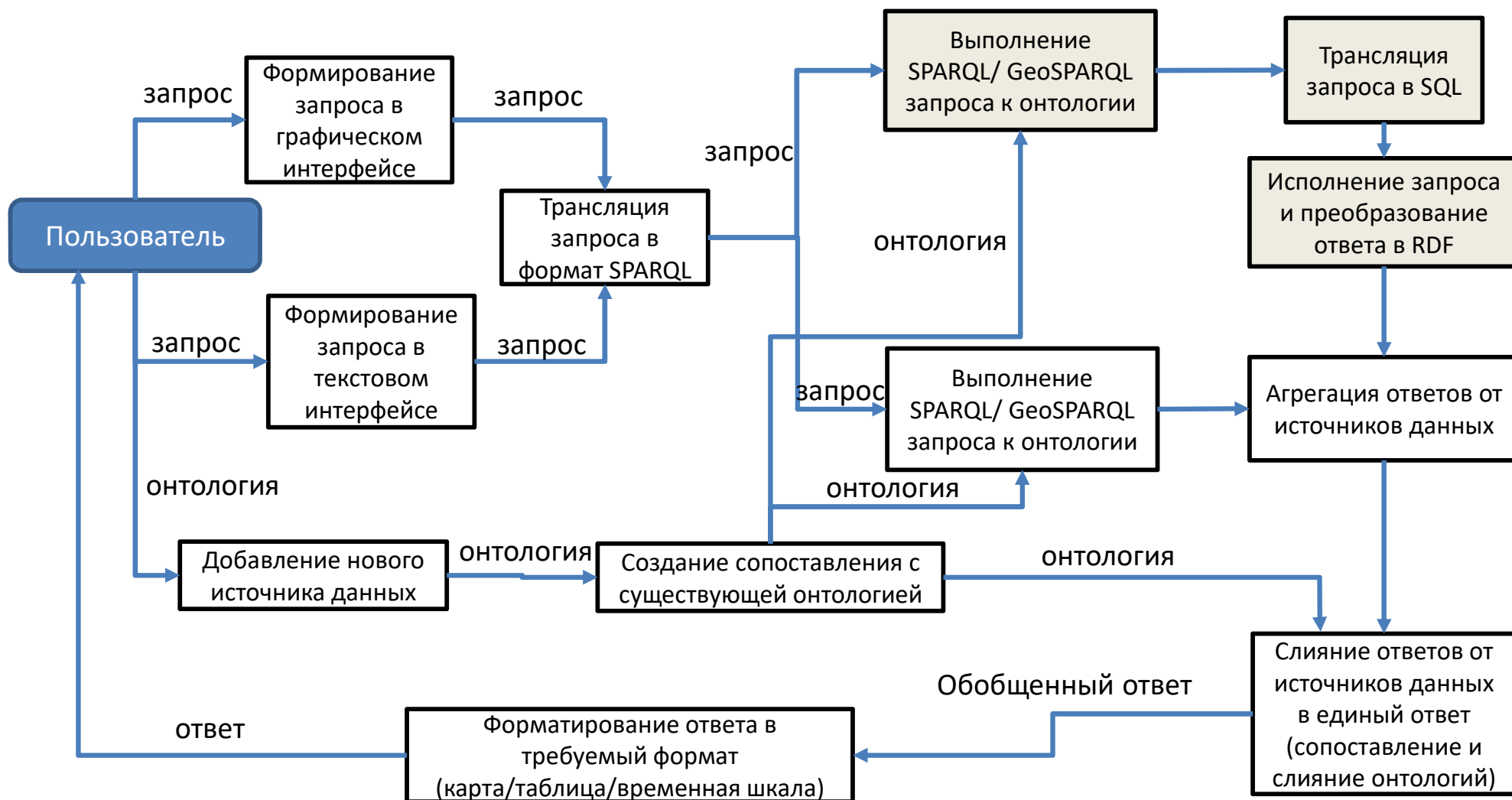


Архитектура платформы для запроса к разнородным источникам данных





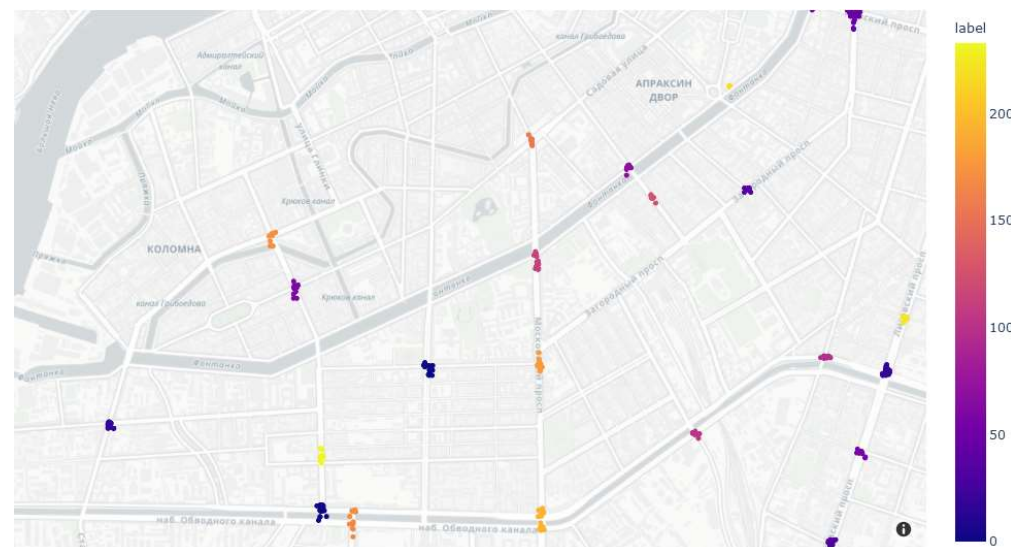
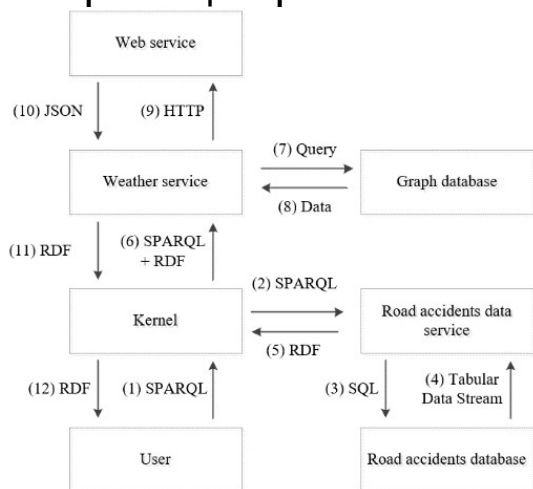
Функциональная модель системы





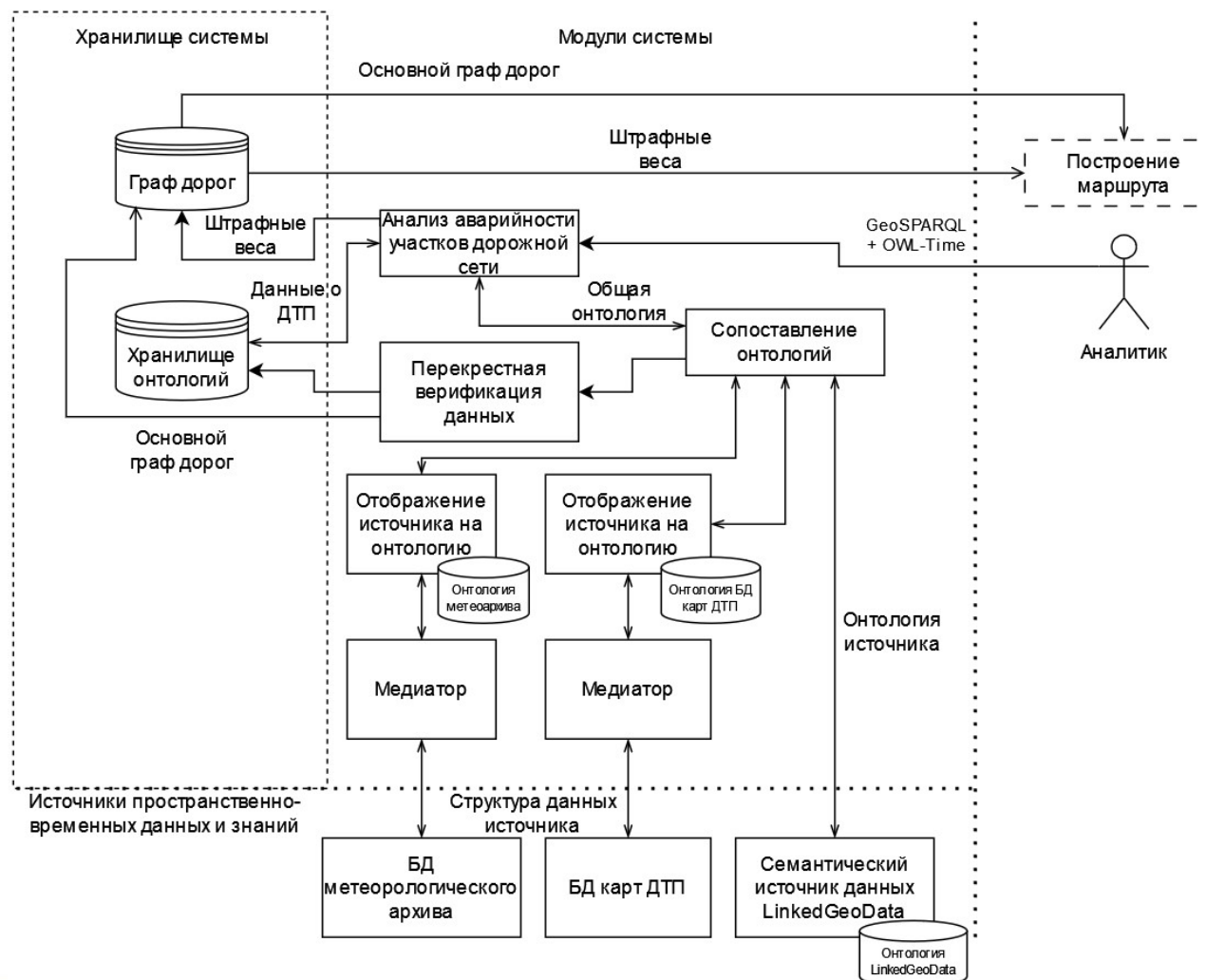
Пример поиска и обработки информации с учетом пространственно-временных отношений

- Объединение разнородных источников данных на примере поиска мест концентрации ДТП и верификации данных карточек ДТП
- Источники данных:
 - Портал статистики ДТП Госавтоинспекции: stat.gibdd.ru (P)
 - LinkedGeoData (C)
 - GeoNames/Nominatim (C)
 - Weather (P)
 - Данные о границах региона



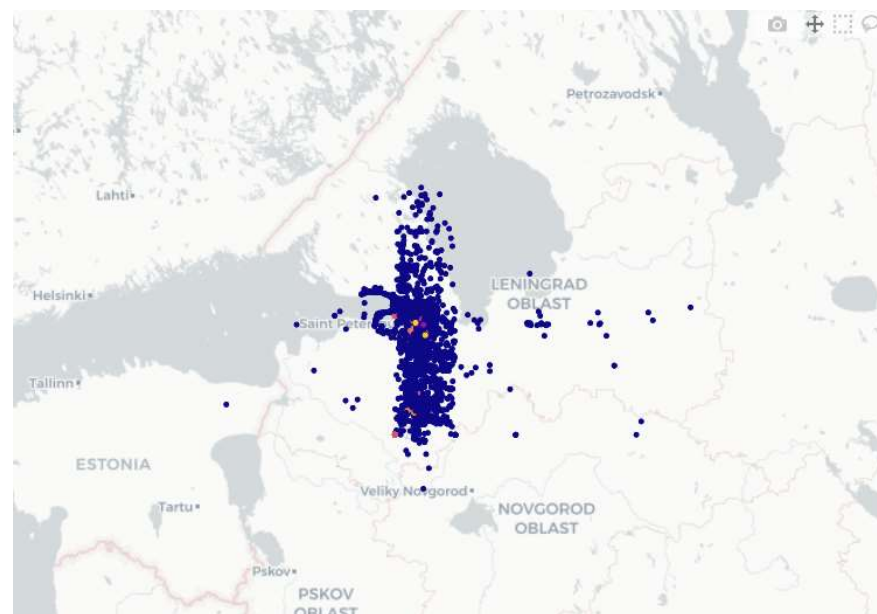
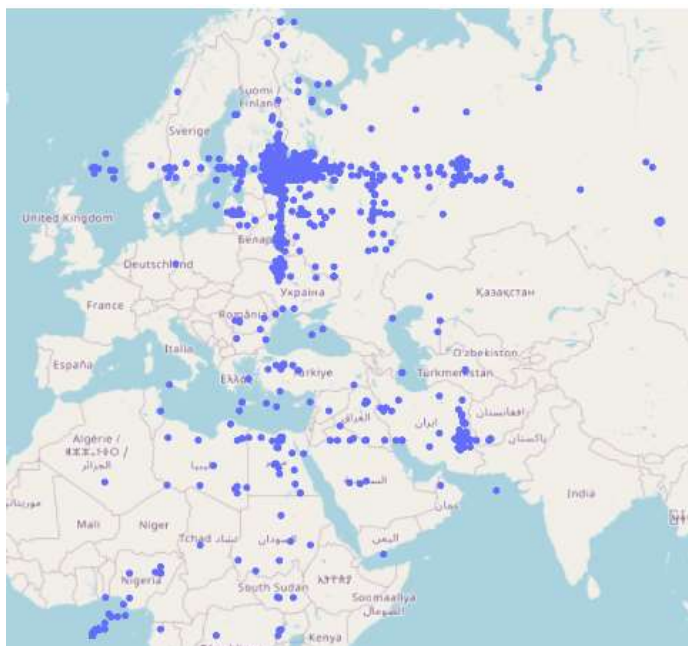


Пример модели системы для поиска информации, имеющей привязку к пространству и времени





Пример верификации данных. Анализ карточек ДТП по СПб и ЛО



Карточка ДТП:

- Координаты
- Адрес
- Тип дороги

LinkedGeoData

Данные о границе регионов

Nominatim

TODO: Замена
геокодера.



Пример R2RML описания реляционного источника данных

[PrefixDeclaration]

rdf : http://www.w3.org/1999/02/22-rdf-syntax-ns#

rdfs : http://www.w3.org/2000/01/rdf-schema#

owl : http://www.w3.org/2002/07/owl#

xsd : http://www.w3.org/2001/XMLSchema#

obda : https://w3id.org/obda/vocabulary#

[MappingDeclaration] @collection [[

mappingId MAPPING - ID1

target _:ontop -bnode -1{ accident_id }/{weather_id } a <http://example.edu/smart-city/stat_gibdd_accident_weather >; <http://example.edu/smart-city/stat_gibdd_accident_weather # accident_id >

{ accident_id }^^xsd:string; <http://example.edu/smart-city/stat_gibdd_accident_weather # weather_id > { weather_id }^^xsd:string.

source SELECT * FROM "public".

stat_gibdd_accident_weather "

mappingId MAPPING - ID2

target _:ontop -bnode -1{

stat_gibdd_accident_weather_accident_id }/{

stat_gibdd_accident_weather_weather_id } <http://example.edu/smart-city/stat_gibdd_accident_weather #

ref - accident_id > <http://example.edu/smart-city/stat_gibdd_accident_cards /id={ stat_gibdd_accident_cards_id }>.

source SELECT "public".

stat_gibdd_accident_weather "." accident_id " AS "

stat_gibdd_accident_weather_accident_id ", "public".

stat_gibdd_accident_weather "." weather_id " AS "

stat_gibdd_accident_weather_weather_id ", "public".

stat_gibdd_accident_cards "." id " AS "

stat_gibdd_accident_cards_id " FROM "public".

stat_gibdd_accident_weather ", "public".

stat_gibdd_accident_cards " WHERE "public".

stat_gibdd_accident_weather "." accident_id " = "public

"." stat_gibdd_accident_cards "." id"

mappingId MAPPING - ID3

target _:ontop -bnode -1{

stat_gibdd_accident_weather_accident_id }/{

stat_gibdd_accident_weather_weather_id } <http://example.edu/smart-city/stat_gibdd_accident_weather #

ref - weather_id > <http://example.edu/smart-city/

stat_gibdd_weathers /id={ stat_gibdd_weathers_id }>.

source SELECT "public".

stat_gibdd_accident_weather "." accident_id " AS "

stat_gibdd_accident_weather_accident_id ", "public".

stat_gibdd_accident_weather "." weather_id " AS "

stat_gibdd_accident_weather_weather_id ", "public".

stat_gibdd_weathers "." id " AS " stat_gibdd_weathers_id

" FROM "public"." stat_gibdd_accident_weather ", "

public"." stat_gibdd_weathers " WHERE "public".

stat_gibdd_accident_weather "." weather_id " = "public

"." stat_gibdd_weathers "." id"



Пример GeoSPARQL к R2RML модели БД карточек ДТП

SPARQL Query

```
# BASE <http://example.edu/smart-city/>
SELECT ?road_importance ?accident_coordinates
WHERE {
  ?cid <http://example.edu/smart-city/stat_gibdd_accident_cards#road_importance> ?id .
  ?iid <http://example.edu/smart-city/stat_gibdd_road_importances#id> ?id .

  ?cid <http://example.edu/smart-city/stat_gibdd_accident_cards#accident_coordinates> ?accident_coordinates .
  ?iid <http://example.edu/smart-city/stat_gibdd_road_importances#road_importance> ?road_importance .
}
```

Show 100 or all results. Use short IRIs

Execution time: 177ms. Solution mappings returned: 100.

SPARQL results SQL translation

road_importance	accident_coordinates
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.300261 59.908539)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.313371 59.909056)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.292547 59.922705)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.307449 59.916016)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.319637 59.927625)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.313168 59.90834)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.317245 59.925883)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.297793 59.909271)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.278707 59.915316)"^^xsd:string
"Местного значения (дорога местного значения, включая относящиеся к собственности поселений, муниципальных районов, г...	"POINT(30.297536 59.912106)"^^xsd:string

street	accident_coordinates
аллея Охтинская	POINT(30.431238 60.047141)
аллея Охтинская	POINT(30.431732 60.046407)
аллея Охтинская	POINT(30.432065 60.046193)
аллея Поликарпова	POINT(30.177397 60.025401)
аллея Поликарпова	POINT(30.289221 60.007601)

Protégé + плагин ontop



Пример федеративного GeoSPARQL запроса к двум источникам

```
BASE <http://example.edu/smart-city/>

SELECT ?id ?accident_coordinates ?weather ?remote_weather

WHERE {
  ?_cid <stat_gibdd_accident_cards#id> ?id .
  ?_awid <stat_gibdd_accident_weather#accident_id> ?id .

  ?_awid <stat_gibdd_accident_weather#weather_id> ?wid .
  ?_wid <stat_gibdd_weathers#id> ?wid .

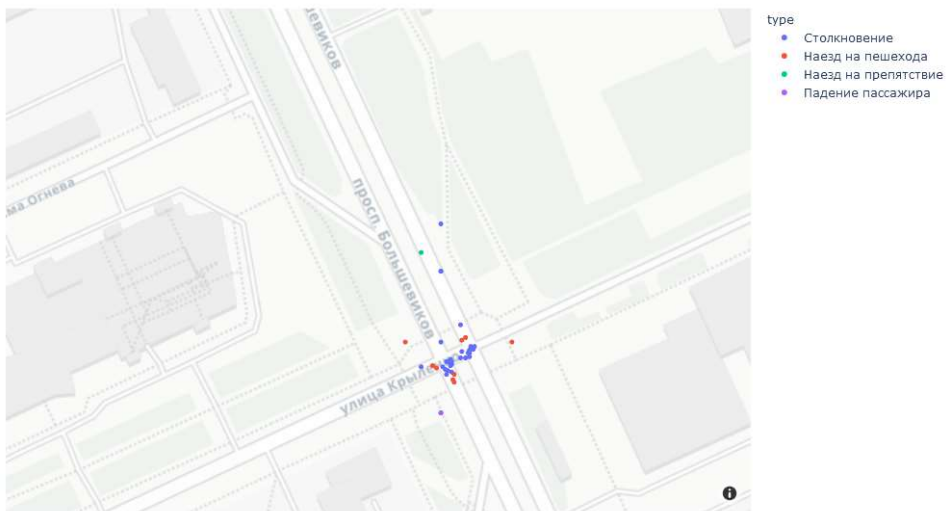
  ?_cid <stat_gibdd_accident_cards#accident_coordinates> ?accident_coordinates .
  ?_wid <stat_gibdd_weathers#weather> ?weather .

  # Подключение внешнего источника.
  # Расширение данных по координатам ДТП.
  #####
  ?_rwid <remote_weather_datasource#coordinates> ?accident_coordinates .
  ?_rwid <remote_weather_datasource#remote_weather> ?remote_weather .
  #####
}
```

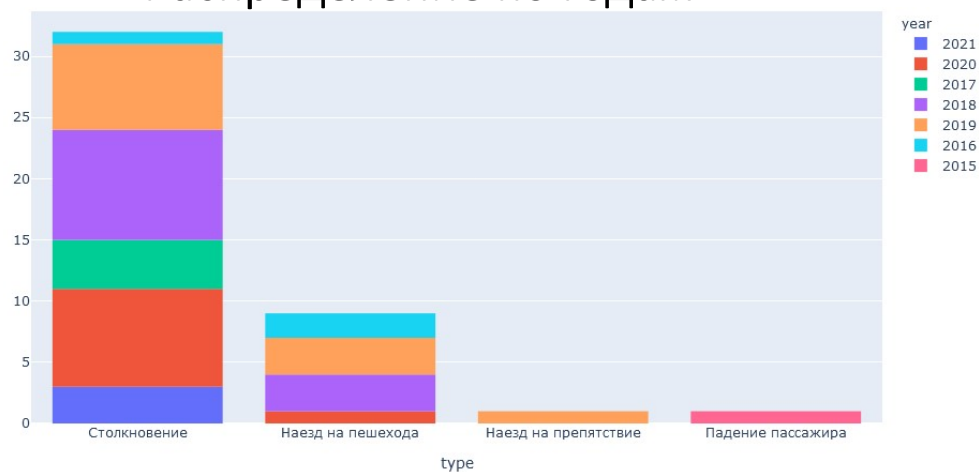



Пример анализа кластера

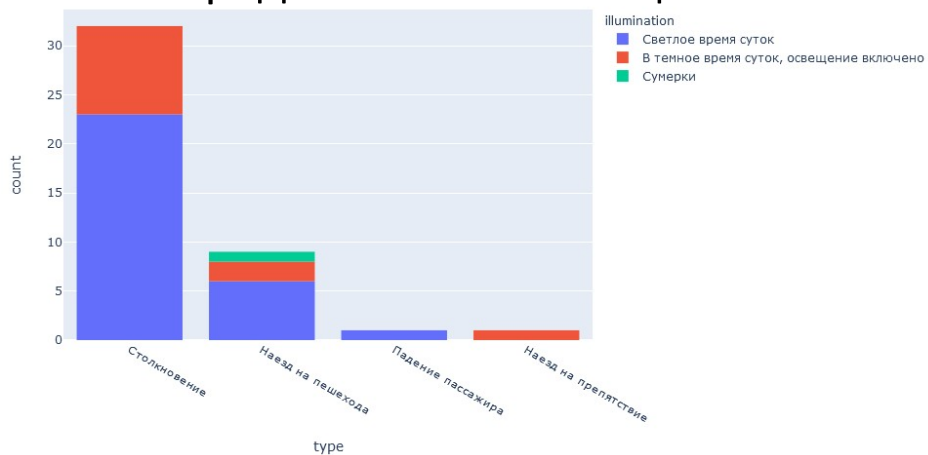
Распределение на карте



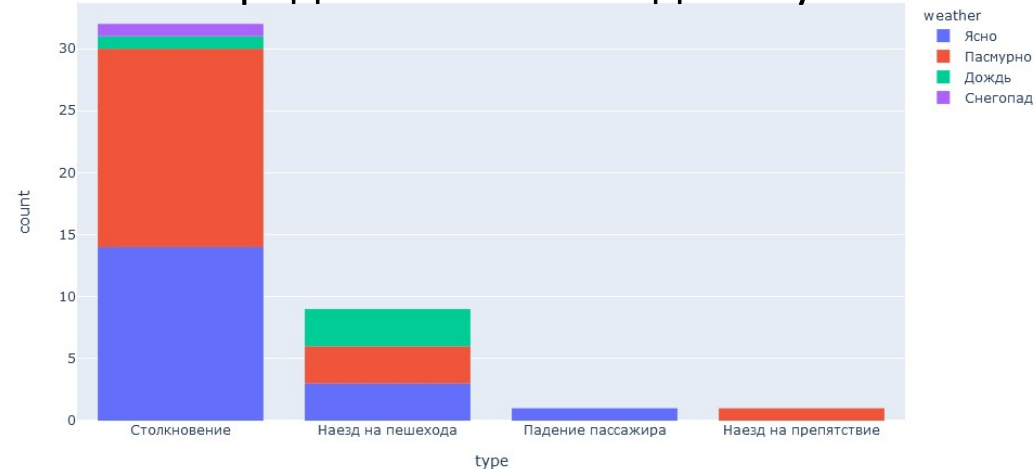
Распределение по годам



Распределение по освещенности



Распределение по погодным условиям





Поддержка исследований

- РФФИ
 - 17-07-00327 Разработка теоретических и технологических основ построения когнитивных ассистентов для поддержки принятия решений при сопоставлении онтологий (2018-2020)
 - 17-07-00328 Разработка метода отображения онтологий, основанного на композиции нейронных сетей (2018-2020)
 - 20-07-00904 Разработка методов и моделей слияния пространственно-временных знаний для интеллектуальной поддержки принятия решений в «умном городе» (2020-2022)
- ИТМО
 - НИР магистрантов и аспирантов «Разработка методов обеспечения информационной поддержки мобильности населения в цифровом пространстве умного города с использованием семантических приложений» (2019)



Публикации

- Volkov, A., Teslya, N., Moskvitin, G., Brovin, N., Bochkarev, E. Spatio-temporal Data Sources Integration with Ontology for Road Accidents Analysis. *Lecture Notes in Business Information Processing*, 2022, 444 LNBIP, pp. 251–262
- Savosin, S., Teslya, N. Estimation and Aggregation Method of Open Data Sources for Road Accident Analysis *Lecture Notes in Networks and Systems*, 2022, 418 LNNS, pp. 1025–1034
- Savosin, S.V., Mikhailov, S.A., Teslya, N.N. Systematization of approaches to assessing the quality of spatio-temporal knowledge sources. *Journal of Physics: Conference Series*, 2021, 1801(1), 012006
- Savosin, S., Teslya, N., Mikhailov, S. Assessment Formation of Open Data Sources during Their Aggregation for Analyzing Road Accidents. *Conference of Open Innovation Association, FRUCT*, 2021, 2021-October, pp. 239–245
- Smirnov, A., Teslya, N. Ontology Matching for Product Lifecycle Management. *IFIP Advances in Information and Communication Technology*, 2020, 594, pp. 256–269
- Teslya N., Ryabchikov I., Mikhailov S. Forming of smart city resident digital identity based on the city sources analysis. 2019 IEEE International Black Sea Conference on Communications and Networking, BlackSeaCom 2019, 7th IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom 2019), Sochi, Russia, 3-6 June 2019, IEEE. 2019. P. 1–3.
- Teslya N., Ryabchikov I., Petrov M., Taramov A., Lipkin E. Smart City Platform Architecture for Citizens' Mobility Support. *Proceedings of the 13th International Symposium "Intelligent Systems 2018"*, Saint Petersburg, 22-24 October 2018, *Procedia Computer Science*, Elsevier. 2019. Vol. 150. P. 646–653.
- Teslya N., Savosin S. Matching Ontologies with Word2Vec-Based Neural Network. *Computational Science and Its Applications – ICCSA 2019, The 19th International Conference on Computational Science and its Applications (ICCSA 2019)*, Saint Petersburg, Russia, 1-4 July 2019, *Lecture Notes in Computer Science*, Springer. 2019. Vol. 11624. P. 745–756.
- Smirnov A., Teslya N., Kashevnik A. Context-aware routing service for intelligent mobile tourist guide. *Proceedings of the 8th International Symposium "Information Fusion and Intelligent Geographical Information Systems 2017"*, Shanghai, China, 10-12 May 2017, Popovich V., Schrenk M., Thill JC., Claramunt C., Wang T. (Ed.), *Lecture Notes in Geoinformation and Cartography*, Springer. 2017. P. 243–258.
- Smirnov A., Teslya N., Savosin S., Shilov N. Ontology Matching for Socio-Cyberphysical Systems: An Approach Based on Background Knowledge. *Internet of Things, Smart Spaces, and Next Generation Networks and Systems. NEW2AN 2017, ruSMART 2017, NsCC 2017, NEW2AN/ruSMART 2017 (ruSmart/New2AN)*, St.Petersburg, Russia, 28-30 August 2017, Galinina O., Andreev S., Balandin S., Koucheryavy Y. (Ed.), *Lecture Notes in Computer Science*, Springer. 2017. Vol. 10531. P. 29–39.



Направления дальнейших исследований

- Автоматизация сопоставления семантики концептов в различных источниках данных.
- Реализация трансляции пространственных и временных отношений в онтологию из реляционных моделей.
- Учет пространственно-временных отношений и концептов онтологии при сопоставлении.
- Использование embedding и нейросетевого подхода для формирования пространств «смыслов» по которым осуществляется поиск сходства концептов.



СПб
ФИЦ
РАН

Санкт-Петербургский
Федеральный исследовательский центр
Российской академии наук



Спасибо за внимание!