Казахский Национальный Университет имени аль-Фараби

УДК 004.043(043.3)

ОСПАН ӘСЕЛ ҒАЛЫМЖАНҚЫЗЫ

Разработка геоинформационных моделей с использованием онтологического подхода для извлечения знаний из гетерогенных источников

8D06105 - Системная инженерия»

Диссертация на соискание степени
доктора философии (PhD)

Научные консультанты:
Заведующая кафедрой искусственного интеллекта и Big Data,
профессор
(КазНУ)
Мансурова Мадина Есимхановна,
Профессор, PhD
Стефан Янг,
(Государственный Университет Салем, США)

Республика Казахстан

Алматы, 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 6
1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОНТОЛОГИЧЕСКОГО ПОДХОДА К МОНИТОРИНГУ ВОДНЫХ РЕСУРСОВ 11
1.1 Онтологии в управлении и мониторинге водных ресурсов 11
1.2 Комплексный обзор онтологий в гидрологии для разработки приложений 18
1.3 Методы семантического вывода и извлечения новых знаний в гидрологических онтологиях 26
2 СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ТАБЛИЧНЫХ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ И ПОПОЛНЕНИЯ ОНТОЛОГИЙ В МОНИТОРИНГЕ ВОДНЫХ РЕСУРСОВ 36
2.1 Типология и формальная модель таблиц в задаче семантической интерпретации 38
2.1.1 Подходы к семантической интерпретации таблиц 44
2.1.2 Методы оценки производительности систем STI 49
2.2 Сбор данных из гетерогенных источников по водному домену Казахстана 52
2.2.1 Область исследования 52
2.2.2 Гетерогенные источники по водному домену Казахстана 55
2.3 Инструменты извлечения таблиц из неструктурированных источников 60
3 МЕТОДОЛОГИЯ ФОРМИРОВАНИЯ ГЕОИНФОРМАЦИОННОЙ МОДЕЛИ НА ОСНОВЕ ОНТОЛОГИЙ ДЛЯ ИЗВЛЕЧЕНИЯ ЗНАНИЙ 64
3.1 Многоуровневая архитектура онтологически-ориентированной геоинформационной модели для извлечения знаний 65
3.2 Модульная архитектура гидроонтологии: выравнивание с международными онтологиями и расширение национальными модулями 71
3.2.1 Сравнение стандартных онтологий и разработанной гидроонтологической модели 73
3.2.2 Модуль наблюдений и качества воды (SOSA/SSN/WHOW) 74
3.2.3 Модуль гидрографических объектов (HY_Features) 76
3.2.4 Национальный модуль нормативов качества воды – WR_KZ 78
3.3 Метод итеративной семантической интерпретации таблиц для пополнения и обновления графов знаний 83
3.4 Применение языка правил семантической сети для автоматического вывода знаний 88
4 РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ 92
4.1 Формирование эталонного набора триплетов для эксперимента 92
4.2 Экспериментальная валидация метода семантической интерпретации таблиц на основе эталонного набора триплетов 94
4.3 Сравнительный анализ предложенного метода STI с существующими методами семантической интерпретации таблиц 97
4.4 Семантическая интеграция и логический вывод в GraphDB 103
4.5 Извлечение и аналитическая эксплуатация данных на основе SPARQL-запросов 117
ЗАКЛЮЧЕНИЕ 123

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 125
ПРИЛОЖЕНИЕ А – Авторское свидетельство 134
ПРИЛОЖЕНИЕ Б – Список публикаций 135

НОРМАТИВНЫЕ ССЫЛКИ

В данной диссертации использовались ссылки согласно следующим стандартам:

ГОСО РК 5.04.034-2011 «Государственный общеобязательный стандарт образования Республики Казахстан. Послевузовское образование. Докторантура». Основные правила заверены Министерством Образования и Науки РК. «17» июнь 2011 ж. №261. Астана 2011.

«Инструкция по оформлению диссертационных работ и авторефератов, МОН РК, Комитет высшей аттестации, Алматы, 2004, МЕСТ 7.1-2003. Библиографическая опись.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ИББ – Иле-Балхашский бассейн
ИЗВ – Индекс загрязнения воды
ИУВР – Интегрированное управление водными ресурсами
ABox – Assertional Box (фактический уровень онтологии)
TBox – Terminological Box (уровень понятий и классов онтологии)
CSV – Comma-Separated Values
DBpedia – база знаний на основе Википедии
F1-score – гармоническое среднее Precision и Recall
GIS – Geographic Information System
GraphDB – графовая RDF-база данных
HY_Features (Hy_Fo) – Hydro Features Ontology
IBB – Ile-Balkhash Basin
JSON – JavaScript Object Notation
OD – Observed Data
OO – Observed Objects
OWL – Web Ontology Language
PDF – Portable Document Format
Precision – точность извлечения
Recall – полнота извлечения
RDF – Resource Description Framework
RDFS – RDF Schema
SPARQL – Query Language for RDF
SQL – Structured Query Language
STI – Semantic Table Interpretation
SOSA – Sensor, Observation, Sample, and Actuator Ontology
SSN – Semantic Sensor Network Ontology
SWRL – Semantic Web Rule Language
TO – Time Ontology
TTL – Turtle (RDF serialization format)
WC – Water Class
WHOW – Water Health Open Knowledge Ontology
WO – Water Objects
WPI – Water Pollution Index
WR – Water Regulations
WR_KZ – Water Regulations Kazakhstan Module
XML – Extensible Markup Language

ВВЕДЕНИЕ

Диссертационная работа посвящена разработке онтологически-ориентированной геоинформационной модели, предназначенной для интеграции, формализации и аналитической обработки гетерогенных данных о водных ресурсах Республики Казахстан. Актуальность исследования обусловлена фрагментарностью хранения гидрологических, гидрохимических, нормативных и социально-экономических данных, представленных в различных форматах и ведомственных источниках, что затрудняет их комплексный анализ и использование в системах мониторинга и принятия решений.

Научная новизна работы заключается в разработке итеративного метода Semantic Table Interpretation (STI) с расширенным механизмом распознавания сущностей, обеспечивающего автоматическое выявление столбцов-субъектов, определение предикатных связей и атрибутивных колонок с последующим формированием RDF-триплетов. Метод сочетает структурные и лексические эвристики (анализ уникальности значений, полноты данных, контекстной согласованности заголовков) с метриками семантической близости, включая нормализованное расстояние Левенштейна и косинусное сходство. Такой подход обеспечивает устойчивость к гетерогенности источников и минимизирует необходимость ручной настройки правил для каждого отдельного формата таблиц.

В рамках исследования разработана модульная гидро-онтологическая модель, выровненная с международными стандартами (SOSA/SSN, HY_Features, WHOW, Time Ontology) и расширенная авторскими национальными модулями WR_KZ (нормативы качества воды Республики Казахстан) и Socio_Geo_KZ (социально-географические характеристики регионов). Онтология объединяет данные наблюдений, химические индикаторы, гидропосты, нормативные пороги и региональные параметры в единую формализованную семантическую структуру.

Для автоматизации вывода знаний реализован механизм правил на основе SWRL, обеспечивающий классификацию качества воды по национальным нормативам, связывание наблюдений с пространственно-временным контекстом и формирование новых логически обоснованных утверждений. Таким образом, граф знаний выполняет не только функцию хранения данных, но и функцию дедуктивной интерпретации предметной области.

Практическая реализация включает воспроизводимый конвейер обработки данных: «PDF/веб/Excel → Таблица → STI → RDF → GraphDB → SPARQL → Геовизуализация». Валидация метода на эталонном наборе продемонстрировала высокие показатели качества формирования триплетов и превосходство над существующими решениями STI.

Результатом работы является создание масштабируемого графа знаний по поверхностным водам Казахстана, обеспечивающего логическую связность, интероперабельность и аналитическую эксплуатацию данных. Разработанная модель формирует основу для построения интеллектуальных геоинформационных систем мониторинга и поддержки принятия решений в

сфере управления водными ресурсами и экологической безопасности.

Ключевые слова: ОНТОЛОГИЯ, ГИДРОЛОГИЯ, СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ТАБЛИЦ, ИЗВЛЕЧЕНИЕ ЗНАНИЙ, ГРАФ ЗНАНИЙ, RDF-ТРИПЛЕТ, ЭВРИСТИЧЕСКИЙ МЕТОД, СЕМАНТИЧЕСКИЕ ПРАВИЛА.

Уровень изучения проблемы. Проблематика интерпретации табличных данных активно развивается в рамках направлений Semantic Web и Knowledge Graph Engineering. Существующие исследования включают эвристические методы (TableMiner, TableMiner++), гибридные и обучаемые конвейеры (QTLTableMiner++, MantisTable, JenTab, MTab), а также подходы на основе глубоких моделей и представлений графов знаний (TURL, DODUO), решающие задачи аннотации ячеек, типов столбцов и связей между ними.

Несмотря на прогресс, большинство подходов ограничиваются задачами маппинга таблиц в RDF и не охватывают полный цикл обработки данных, включающий онтологическое выравнивание, интеграцию гетерогенных источников, формализацию нормативов и правило-ориентированный вывод. В водной сфере международные онтологии недостаточно адаптированы к национальным нормативным требованиям и социально-географическому контексту.

Остаются нерешёнными вопросы устойчивости STI-систем к гетерогенности источников и формализации нормативных вычислений в машиночитаемом виде. В связи с этим построение онтологически-ориентированной геоинформационной модели водного домена, обеспечивающей интеграцию данных и автоматический вывод знаний, остаётся актуальной научной задачей.

Цель диссертационной работы. Разработать онтологически-ориентированную геоинформационную модель для установления связей между объектами водного домена Казахстана на основе данных из гетерогенных источников.

Задачи исследования:

1. Спроектировать архитектуру геоинформационной модели с онтологическим подходом для извлечения знаний.
2. Разработать итеративный подход семантической интерпретации таблиц с расширенным распознаванием сущностей.
3. Построить гидро-онтологическую модель, обеспечивающую выравнивание с международными онтологиями и расширяемую новыми модулями.
4. Расширить семантику онтологии и автоматизировать вывод знаний на основе правил Semantic Web Rule Language (SWRL).

Объектом исследования являются данные из гетерогенных источников о поверхностных вод Иле-Балхашского бассейна.

Методы исследования – онтологическое моделирование и выравнивание, методы семантической интерпретации таблиц, правило-ориентированный вывод из графов знаний

Предметы исследования - методы семантической интерпретации

табличных данных и онтологическое моделирование.

Научная новизна:

1. Предложен улучшенный итеративный подход Semantic Table Interpretation (STI), который превосходит существующие методы (TableMiner++, QTLMiner++, Mantis Table) за счёт:

1.1. применения расширенных эвристических правил для более точного обнаружения столбцов-субъектов и предикатов;

1.2. внедрения механизма автоматического пополнения и обновления графов знаний на основе извлечённых RDF-триплетов.

2. Разработана новая гидро-онтологическая модель Казахстана, выровненная с международными онтологиями (WHOW, Hy_Features) и дополненная:

2.1. новым модулем Water_Regulations_KZ для классификации вод по нормативам Водного кодекса РК;

2.2. модулем Socio_Geo_KZ, связывающим данные о водных объектах с социально-экономическими показателями регионов.

Положения выносимые на защиту:

1. Итеративный подход Semantic Table Interpretation (STI) с расширенными эвристиками для детекции столбца-субъекта и предикатов, обеспечивающий более точный маппинг таблиц в RDF и превосходящий TableMiner++, QTLTableMiner++ и MantisTable по метрикам F1, с встроенным механизмом автоматического пополнения и обновления графов знаний.

2. Гидро-онтологическая модель Казахстана, выровненная с международными стандартами (WHOW, HY_Features) и расширенная модулями Water_Regulations_KZ и Socio_Geo_KZ, с поддержкой правило-ориентированного вывода (SWRL) для автоматического присвоения классов воды и выявления связей «водные объекты ↔ нормы ↔ наблюдения ↔ регионы».

Теоретическая значимость работы состоит в формализации итеративного процесса STI как последовательности детекции ролей столбцов, выявления предикатных кандидатов и атрибутных колонок с последующим онтологическим выравниванием и правиловым выводом. Предложена модульная гидроонтология, объединяющая международные схемы наблюдений/гидрографических объектов и национальные доменные знания (WR_KZ, SocioGeo_KZ). Введён набор SWRL-правил, который эксплицирует нормативно-правовые зависимости (пороговые классификаторы качества воды) и причинно-следственные связи между наблюдениями, географией и социальными факторами.

Практическая значимость заключается в создании воспроизводимого конвейера «Таблица → RDF → GraphDB → SPARQL → карта», обеспечивающего автоматизированное пополнение графа знаний по водным ресурсам Казахстана из PDF, веб-страниц и Excel-файлов. Система избавляет от ручной перекодировки таблиц, снижает стоимость интеграции новых источников и даёт инструмент для оперативного мониторинга: от обнаружения

превышений по индикаторам до анализа пространственно-временных закономерностей в разрезе бассейнов и регионов. Результаты могут использоваться госорганами и водохозяйственными организациями для отчётности, планирования мероприятий по качеству воды и оценки рисков. Архитектура модели открыта к расширению: добавление новых модулей (гидротехнические сооружения, источники загрязнения) и правил вывода обеспечивает масштабирование к национальному уровню.

Достоверность и обоснованность полученных результатов. Достоверность полученных результатов обеспечивается использованием формализованных стандартов Semantic Web, выравниванием разработанной онтологической модели с международными онтологиями и экспериментальным сравнением предложенного STI-подхода с существующими методами, показавшим его превосходство по метрикам качества. Обоснованность результатов подтверждается апробацией модели на реальных данных Иле-Балхашского бассейна, логической консистентностью сформированного графа знаний и публикацией основных положений исследования в рецензируемых научных изданиях.

Связь темы с планами научно-исследовательских программ. Диссертационная работа выполнена в рамках научно-исследовательских проектовМОН РК (источник финансирования Комитет науки МОН РК):

- грантового финансирования (ГФ) AP09261344 «Разработка методов автоматического извлечения геопространственных объектов из гетерогенных источников для информационного обеспечения геоинформационных систем» в 2021–2023 годы;

- программно-целевого финансирования (ПЦФ) BR24993001 «Создание большой языковой модели (LLM) для поддержки казахского языка и технологического прогресса» в 2024–2026 годы.

Апробация работы. Основные результаты диссертационной работы доложены и обсуждены на международных и зарубежных научных конференциях, научных семинарах:

– AIES 2022: International Conference Artificial Intelligence in Engineering and Science (Швейцария 2022);

– Научном семинаре кафедры информатики факультета информационных технологий КазНУ имени аль-Фараби.

Структура и краткое содержание работы. Диссертация состоит из введения, четырех разделов, заключения и списка использованной литературы.

Во введении обоснована актуальность темы, сформулированы цель и задачи исследования, определены объект, предмет и методы исследования, раскрыты научная новизна, теоретическая и практическая значимость работы.

В первом разделе представлены теоретические основы онтологического подхода к мониторингу водных ресурсов, выполнен анализ существующих онтологий в гидрологии, а также рассмотрены методы семантического вывода и извлечения новых знаний в гидрологических моделях.

Во втором разделе исследованы подходы к семантической интерпретации

табличных данных, методы оценки эффективности STI-систем, проведён сбор и анализ гетерогенных источников по водному домену Казахстана, а также рассмотрены инструменты извлечения таблиц из неструктурированных источников.

В третьем разделе разработана методология формирования онтологически-ориентированной геоинформационной модели, представлена её многоуровневая и модульная архитектура, описаны модули наблюдений, гидрографических объектов и национальных нормативов, предложен итеративный метод STI и механизм правило-ориентированного вывода на основе SWRL.

В четвертом разделе приведены результаты экспериментальной валидации разработанного подхода, выполнен сравнительный анализ с существующими методами семантической интерпретации таблиц, представлены результаты формирования новых аксиом после применения SWRL-правил и продемонстрирована реализация интегрированной гидролого-социальной онтологии в среде GraphDB.

В заключении сформулированы основные результаты и выводы диссертационного исследования.

Структура и объем дисертации. Общий объем диссертационной работы – 135 страниц, всего 52 рисунка и 17 таблиц. Работа состоит из введения, 4 разделов, заключения, списка использованной литературы.

.

1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ОНТОЛОГИЧЕСКОГО ПОДХОДА К МОНИТОРИНГУ ВОДНЫХ РЕСУРСОВ

1.1 Онтологии в управлении и мониторинге водных ресурсов

Онтология в информатике определяется как явная спецификация разделяемой концептуализации, что подразумевает создание формальной модели предметной области с описанием её основных сущностей, свойств и отношений [1-2]. В контексте управления водными ресурсами она обеспечивает формализацию знаний о гидрологических объектах, процессах и наблюдениях, позволяя не только хранить данные, но и интерпретировать их в семантическом контексте.

Современная цифровая эпоха характеризуется взрывным ростом объёмов данных, генерируемых сенсорными сетями, спутниковыми системами дистанционного зондирования, гидрологическими моделями и симуляциями, социальными сетями и краудсорсинговыми платформами [3]. По прогнозам, к 2025 году мировой объём цифровых данных достигнет 175 зеттабайт [4], однако около 70% этих данных никогда не используется из-за их разрозненности и отсутствия совместимости между изолированными хранилищами («информационными силосами»). Дополнительной проблемой является необходимость использования программных инструментов, способных автоматически считывать, обрабатывать и интегрировать данные из гетерогенных источников.

В науках о Земле и гидрологии особенно актуальна проблема гетерогенности данных. Термины, применяемые для описания наблюдений, различаются, а сами данные часто имеют неструктурированный, неполный и разнообразный характер, что усложняет их согласование и интеграцию [5-6]. Для эффективного управления, анализа и передачи таких данных необходимы современные инструменты семантической интеграции.

Одним из наиболее эффективных решений является использование онтологий и графов знаний [7]. Онтология формализует предметную область, определяя классы, свойства и отношения между сущностями, а также правила их взаимодействия [8]. Она служит концептуальной основой для семантического описания данных и обеспечивает возможность логического вывода новых знаний. Граф знаний является конкретной реализацией онтологии, наполненной фактами: узлы отражают конкретные объекты, а рёбра фиксируют их взаимосвязи [9]. Если онтология задаёт структуру и правила организации знаний, то граф знаний позволяет визуализировать и анализировать фактические данные, объединяя разнородные источники в единую семантическую сеть.

В области гидрологии и мониторинга водных ресурсов онтологический подход зарекомендовал себя как эффективный инструмент решения следующих задач:

- интеграция данных из сенсорных сетей, спутников, лабораторных анализов и нормативно-правовых документов;
- стандартизация терминов и обеспечение совместимости между системами

мониторинга;

- поддержка систем поддержки принятия решений (DSS) и прогнозных моделей;
- выявление пространственно-временных взаимосвязей, критически важных для оценки качества воды и управления ресурсами.

В последние два десятилетия в рамках международного научного сообщества сформирован ряд онтологий и стандартов, направленных на решение задач интеграции, интероперабельности и семантического анализа экологических и гидрологических данных. Эти разработки стали фундаментом для построения интеллектуальных систем поддержки принятия решений (DSS) и геоинформационных платформ нового поколения, способных обрабатывать большие объёмы гетерогенных данных.

В данном исследовании были проанализированы наиболее популярные онтологии, применяемые в управлении и мониторинге водных ресурсов, такие как:

1. SSN/SOSA (Semantic Sensor Network / Sensor, Observation, Sample, and Actuator) [10];
2. OPO (Observational Process Ontology) [11];
3. HydroOntology [12];
4. DiHydro [13].

Ниже представлена сравнительная таблица основных онтологий, отражающая их предметную область(Таблица 1.1).

Таблица 1.1 – Сравнительная таблица основных онтологий

Онтология	Предметная область	Ключевые классы и сущности	Методология
SSN/SOSA [10]	Сенсорные сети, наблюдения, исполнительные механизмы	Sensor, Observation, Sample, Actuator, FeatureOfInterest, Result	Основана на OWL и RDF; описание сенсорных данных и процессов измерений в виде триплетов; поддержка потоковой интеграции данных; совместимость с OGC SensorThings API.
OPO [11]	Процессы наблюдений и управления качеством воды	Observation Acquisition, Observation Data, Observation Processing, Observation Product, Sensor	Моделирование жизненного цикла наблюдений; поддержка логического вывода для отслеживания происхождения и качества данных.

Продолжение Таблицы 1.1

HydroOntology [12]	Гидрологические процессы и пространственные связи	GroundWater, Underground Water, VadoseZone, UnsaturatedZone, SaturatedZone, ZoneOfAeration	Модульная архитектура; расширение онтологий SWEET; совместимость с GeoSPARQL, WMS, WFS; интеграция данных сенсорных сетей, дистанционного зондирования и гидромоделей.
DiHydro [13]	Децентрализованное управление водными ресурсами (DWM)	Node, NodeInput, NodeOutput, Event, SuggestedAction, CurrentObservation, SoilMoistureObservation, WaterFlowObservation	Двухуровневая архитектура (COL и DOL); расширение SSN доменными понятиями; использование SWRL и событийных правил; поддержка in-network processing для WSN.

После рассмотрения сравнительной характеристики онтологий ниже подробно описываются особенности, методология и эффективность каждой из них, включая принципы построения, примеры классов, использование правил SWRL и интеграцию с геоинформационными системами.

Стандарт SSN/SOSA (Semantic Sensor Network / Sensor, Observation, Sample, and Actuator)**,** разработанный в рамках W3C и OGC, предназначен для семантического описания сенсорных сетей, наблюдений, образцов и исполнительных механизмов [10]. Методология построена на описании данных в виде триплетов субъект–свойство–значение**,** где каждый элемент процесса измерений (сенсор, наблюдение, результат, актор) имеет формализованное описание на основе RDF и OWL**.** Такая структура обеспечивает интероперабельность между разнородными сенсорными системами, поддержку потоковой интеграции данных и совместимость с веб-сервисами по стандартам OGC (например, SensorThings API). Эффективность применения выражается в возможности унификации и стандартизации сенсорных данных**,** снижении затрат на их интеграцию в геоинформационные системы и обеспечении оперативного доступа к информации для задач мониторинга и прогноза паводков, что особенно актуально для реального времени.

Обзор базовой структуры онтологии SSN с акцентом на общие шаблоны, используемые в трёх видах деятельности, с объединением классов, где они играют аналогичную роль показан на Рис.1.1.

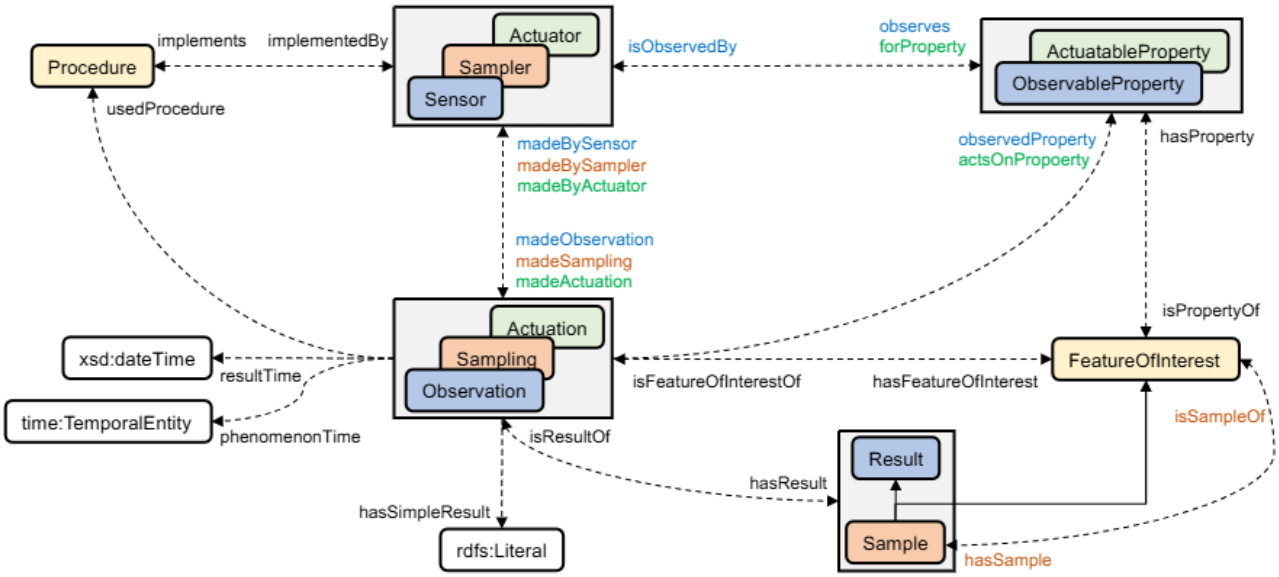


Рисунок 1.1 – Базовая структура онтологии SSN с акцентом на общие шаблоны [10]

Показанные элементы заимствованы из модулей SOSA и SSN, а классы и типы, взятые из внешних словарей, обозначаются префиксом пространства имён и не имеют цветовой заливки. Классы и свойства, выделенные синим цветом, относятся к наблюдениям, красным – к выборке, а зелёным – к срабатыванию (актуированию). Классы, выделенные жёлтым цветом, используются во всех трёх приложениях. Полный набор обратных свойств определён в онтологии, однако на рисунке показано только их подмножество.

Следующий стандарт это - онтология процесса наблюдения OPO (Observational Process Ontology), которая ориентирована на моделирование процессов наблюдения, включая сбор, предварительную обработку, валидацию и интерпретацию данных качества воды [11]. Методология основывается на описании жизненного цикла наблюдений, начиная с планирования измерений и заканчивая публикацией результатов. Основные модули онтологии OPO приведены на Рис.1.2.

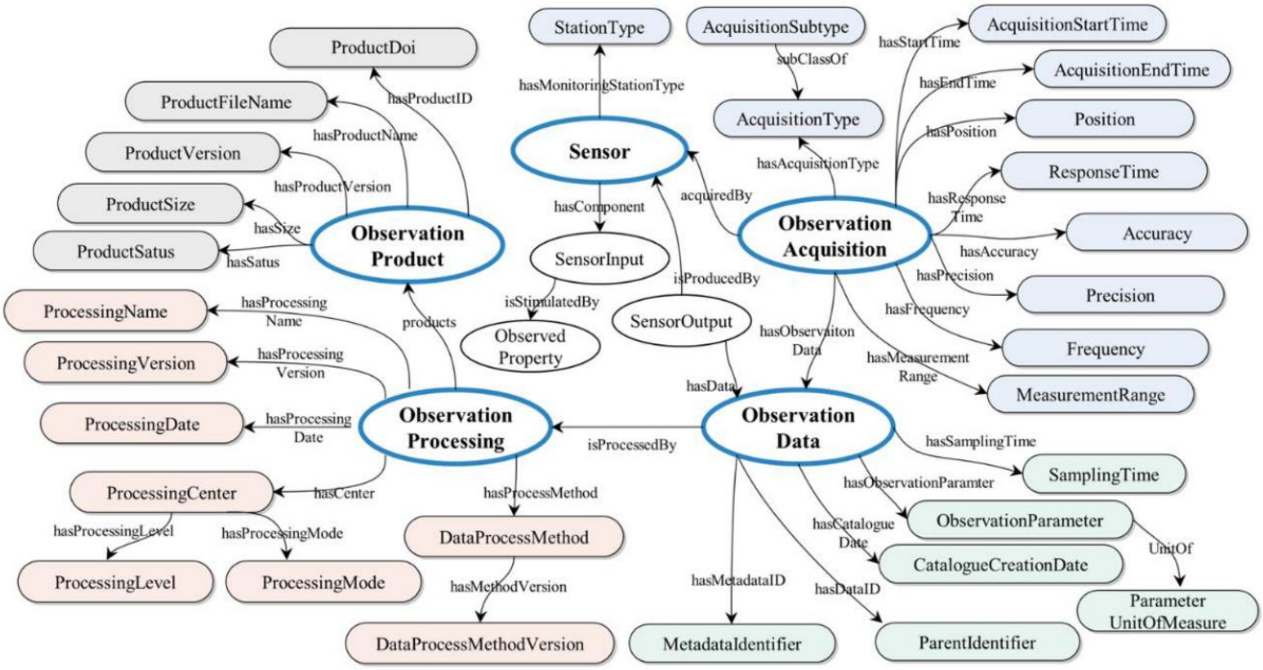


Рисунок 1.2 – Основные модули онтологии процесса наблюдения [11]

Здесь классы и свойства OPO описываются с помощью четырёх основных классов (Получение данных наблюдения, Данные наблюдения, Обработка наблюдений и Продукт наблюдения) и класса Датчик. Он отображает связи, участвующие на каждом этапе процесса наблюдения.

Использование логического вывода позволяет прослеживать происхождение данных, определять причинно-следственные связи между процессами и результатами, а также выявлять ошибки в цепочках обработки**.** Эффективность данной онтологии проявляется в повышении доверия к данным, автоматизации процедур верификации качества и формировании обоснованных аналитических отчётов. В системах поддержки принятия решений, связанных с управлением качеством воды, использование OPO позволяет минимизировать влияние человеческого фактора и обеспечить трассируемость информации на всех этапах.

HydroOntology представляет собой модульную онтологию, разработанную для формализации гидрологических процессов, топологии речных бассейнов и пространственных взаимосвязей между водными объектами [12]. Она была создана как расширение верхнеуровневых онтологий SWEET (Semantic Web for Earth and Environmental Terminology) [14], что обеспечивает её совместимость с другими онтологиями в области наук о Земле. Ее разработка направлена на создание единого семантического пространства для анализа пространственно-временных данных и интеграции разнородных источников информации, включая данные гидропостов и сенсорных сетей, результаты спутникового дистанционного зондирования, гидрологические модели (например, HEC-HMS, SWAT)и геопространственные стандарты (GeoSPARQL, WMS, WFS). Методология HydroOntology основана на принципах модульности, расширяемости и совместимости. На Рисунке 1.3. представлена переработанная

схема классов HydroOntology, иллюстрирующая взаимосвязи между GroundWater, UndergroundWater, VadoseZone, UnsaturatedZone, ZoneOfAeration и SaturatedZone. Показано, что GroundWater теперь локализуется исключительно в насыщенной зоне, а UndergroundWater охватывает как ненасыщенную, так и насыщенную зоны. Такая реконструкция обеспечивает более точное соответствие гидрогеологическим определениям и совместимость с верхнеуровневыми онтологиями SWEET.

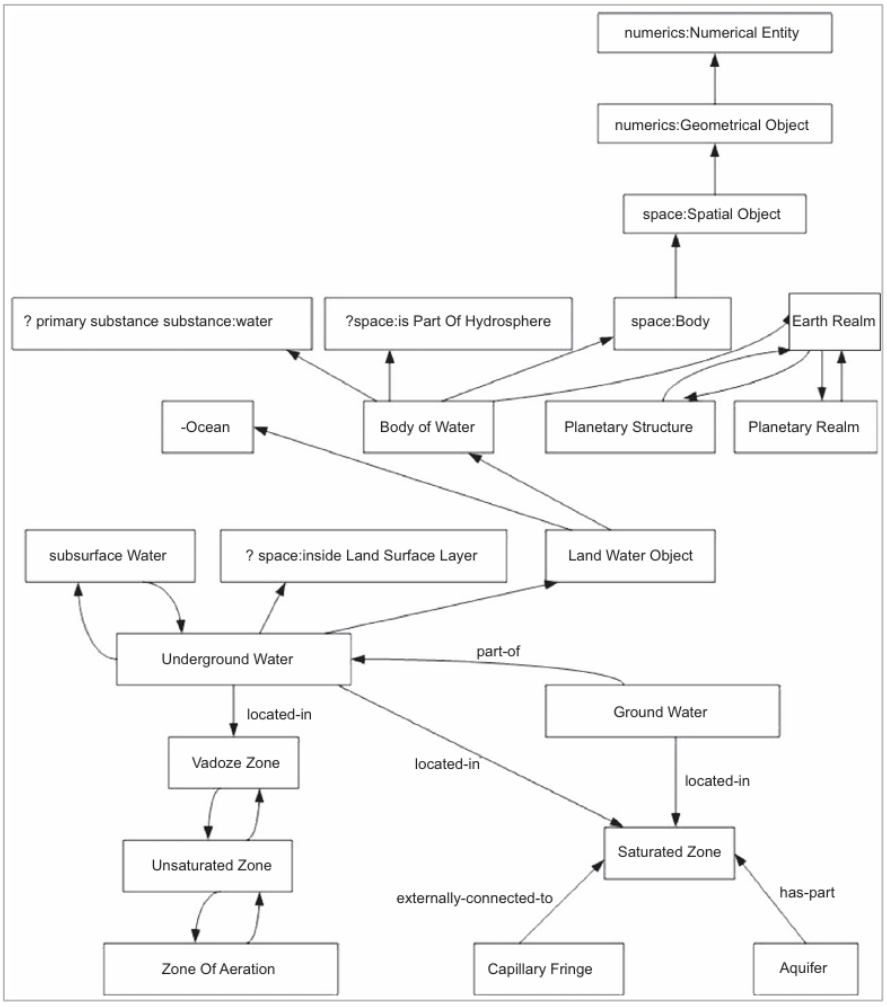


Рисунок 1.3 – Модифицированная схема классов подземных вод в HydroOntology на основе SWEET [12]

Онтологический метод DiHydro [13] реализует онтологический подход к децентрализованному управлению водными ресурсами (Decentralized Water Management, DWM) с использованием беспроводных сенсорных сетей (WSN). Фреймворк опирается на двухуровневую архитектуру онтологии, включающую:

1. Core Ontology Layer (COL) – ядро онтологии, формализующее низкоуровневые данные от сенсоров (расход воды, влажность почвы, проводимость и температура воды);
2. Deductive Ontology Layer (DOL) – уровень дедукции, в котором определяются абстрактные события и выводятся высокоуровневые знания, например garden_needs_irrigation или tank_needs_topping_up.

Методология расширяет стандартную SSN-онтологию (Semantic Sensor Network) доменными понятиями: источники и стоки воды (grey water, green water, black water), сенсорные узлы (Node), актуаторы (Actuator), события (Event) и предлагаемые действия (Suggested Actions). Логический вывод реализуется через правила SWRL и шаблоны событий, что позволяет автоматически формировать рекомендации для управления, включая полив, подпитку баков и предотвращение утечек.

Эффективность метода заключается в том, что локальная обработка данных в узлах WSN снижает энергопотребление и задержки передачи, а онтологическая интеграция обеспечивает семантическую совместимость данных и возможность интеллектуального анализа. DiHydro поддерживает реактивное и событийно-ориентированное управление, способное автоматически выявлять отклонения, оптимизировать использование серой воды и формировать меры реагирования без участия оператора.

На рисунке 1.4 представлена концептуальная модель класса Observation, используемого для описания результатов работы сенсорных узлов в DiHydro

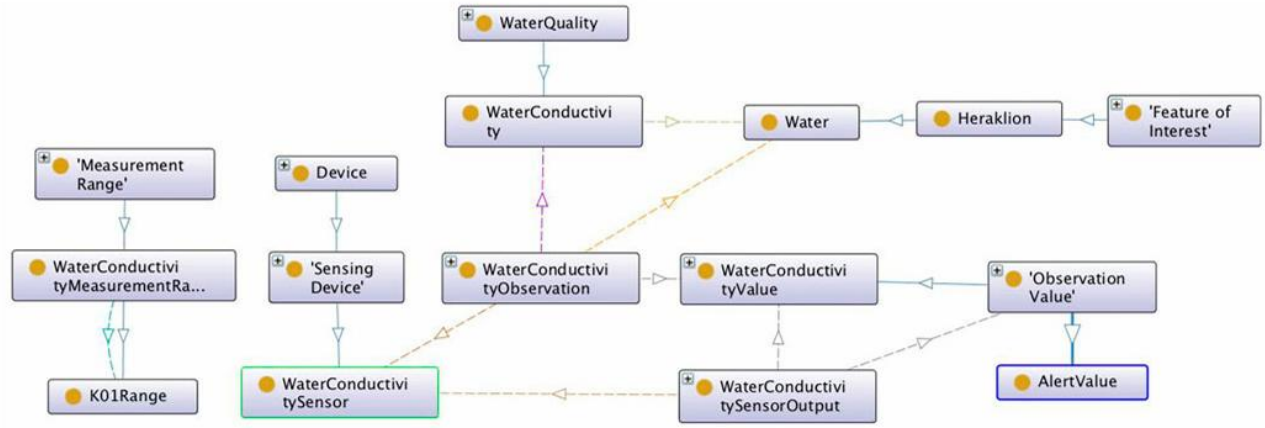


Рисунок 1.4 – Структура сущности «Observation» (наблюдение) в онтологии DiHydro [13]

Здесь для каждого типа сенсора создаётся соответствующий тип наблюдения, такие как: CurrentObservation – наблюдение электрического тока; SoilMoistureObservation – наблюдение влажности почвы; WaterFlowObservation – наблюдение расхода воды.

Все наблюдения DiHydro являются подклассами SSN:Observation и содержат три ключевых элемента:

1. Значение наблюдения (Observation value), представленное в виде диапазона значений (SSN:Region) с учётом погрешности сенсора;
2. Временную метку (timestamp), фиксирующую момент получения данных;
3. Оценку качества (quality estimate), отражающую достоверность измерения.

Такая структура обеспечивает семантическую совместимость с SSN-онтологиями и поддерживает дальнейший логический вывод для управления децентрализованными водными системами.

Таким образом, DiHydro формирует основу для интеллектуальных систем поддержки принятия решений (DSS) в области управления водой, объединяя сенсорные наблюдения, семантическое моделирование и автоматизированный логический вывод.

Внедрение онтологий позволяет создавать интеллектуальные системы поддержки принятия решений (DSS), которые повышают точность мониторинга и прогноза водных процессов, а также оптимизируют управление водными ресурсами. Разработанные международные стандарты и методологии демонстрируют эффективность в автоматизации процессов, сокращении влияния человеческого фактора и формировании доверенных аналитических отчетов. Таким образом, онтологии формируют основу цифровой трансформации гидрологии и служат базисом для построения современных геоинформационных платформ и сервисов экологического мониторинга.

1.2 Комплексный обзор онтологий в гидрологии для разработки приложений

В гидрологии, как и в других сферах окружающей среды, существует огромное количество разнообразных данных из гетерогенных источников, которые не обладают необходимой совместимостью и связностью. Следует отметить, что в то время, как в других областях экологических наук ведутся активные исследования по разработке онтологий, в области гидрологии и управления водными ресурсами их число остается минимальным. Эффективное управление данными и информацией необходимо для понимания и обработки этих сложных структур данных, а также для перспективных оценок и планирования на их основе. Для этого крайне важно развитие онтологий в гидрологии, направленное на улучшение и автоматизацию управления и обработки данных, а также на создание основы для следующего поколения интеллектуальных информационных и систем поддержки принятия решений [15].

Традиционные системы обмена данными и информацией ограничены в способности интегрировать удалённые источники, осуществлять визуализацию, анализировать и передавать результаты моделирования [16]. Поскольку они обычно ориентированы на вычислительные задачи, а не на взаимодействие с пользователями, такие системы зачастую предоставляют только статические отчёты или карты, при этом взаимодействие с данными ограничено. В противоположность этому, информационные системы включают обладающие возможностями веб-управления, визуализации и обмена данными для временных рядов данных окружающей среды с использованием веб-сервисов [17].

Открытая сеть знаний (OKN - Open Knowledge Network) представляет собой открытую и совместно используемую инфраструктуру для соединения связанных данных и сущностей [18]. OKN включает в себя онтологии, знания графы и семантическую паутину. На рисунке 1.5 показана основная схема OKN.

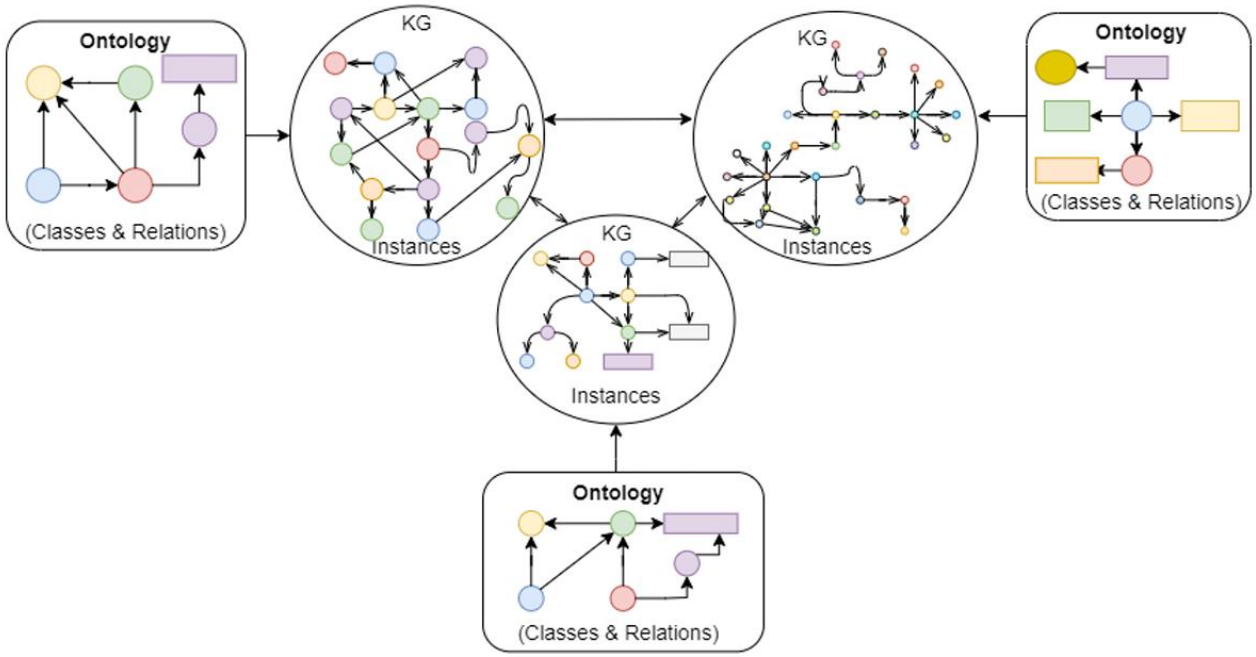


Рисунок 1.5 – Графическое представление сети открытых знаний [19]

Граф знаний (KG – Knowledge Graph) – это многореляционная графовая структура, состоящая из узлов (сущностей) и рёбер (отношений). Семантическая паутина – это расширение современного интернета, в котором информация приобретает четко определённое значение, что облегчает совместную работу машин и человека [19]. Развитие семантических паутин основано на онтологиях, которые являются эффективным инструментом для представления и обмена знаниями [20-24]

Основная задача при создании онтологий – обеспечить общее понимание структуры информации как среди людей, так и среди программных агентов [25]. В период с 2002 по 2025 годы наблюдается значительный рост числа публикаций, связанных с развитием онтологий как в общем, так и в прикладном контексте. На Рисунке 1.6 показаны результаты поиска по базе ScienceDirect (https://www.sciencedirect.com), в котором использовались ключевые слова «Онтология» и «Гидрология»/«Водные ресурсы».



Рисунке 1.6 – Количество публикаций по теме «онтология» и «гидрология» / «водные ресурсы» за 2015–2025 годы по базе ScienceDirect (https://www.sciencedirect.com)

Как видно, здесь наблюдается устойчивый рост количества научных публикаций по обеим тематикам: с 635 и 845 публикаций в 2015 году до 4317 и 6717 публикаций соответственно в 2025 году. Особенно резкий рост заметен после 2020 года, что может свидетельствовать об усилении интереса научного сообщества к интеграции онтологических подходов в области гидрологических исследований и управления водными ресурсами. Наибольшая активность приходится на 2024 год, когда количество публикаций по теме «Онтология» и «Водные ресурсы» достигло максимального значения – 6946. Эти данные подтверждают возрастающую значимость онтологического моделирования в решении задач устойчивого управления водными ресурсами.

В процессе разработки онтологий применяется широкий спектр языков представления знаний, каждый из которых обладает определёнными характеристиками, необходимыми для эффективного моделирования предметной области [26]. Языки онтологий должны соответствовать ряду ключевых требований: наличие чёткого синтаксиса и формальной семантики, поддержка механизмов логического вывода, достаточная выразительная мощность, а также удобство и гибкость в моделировании знаний. Кроме того, критически важным является обеспечение машиночитаемости и интерпретируемости создаваемых онтологий как для программных систем, так и для специалистов в предметной области [27].

Наиболее распространёнными языками представления онтологий являются Web Ontology Language (OWL), а также Resource Description Framework (RDF и его расширение RDF Schema – RDF-S). Эти языки разработаны с учётом

совместимости с ранее существовавшими стандартами, такими как SHOE и DAML+OIL, и адаптированы для задач онтологического моделирования в контексте семантического веба [28]. Язык RDF, разработанный Всемирным консорциумом W3C, основан на использовании XML (Extensible Markup Language) и служит для описания ресурсов в веб-пространстве, обеспечивая универсальное средство структурирования информации [29].

Наряду с языками представления особое значение имеют инструменты визуализации, редактирования и управления онтологиями, которые активно используются как инженерами знаний, так и исследователями. Среди наиболее известных и широко применяемых инструментов можно выделить Protégé – открытое программное обеспечение, поддерживающее OWL и RDF [25]; Hozo – инструмент для семантического проектирования онтологий (http://www.hozo.jp); OntoGraf – модуль для графической визуализации онтологий, интегрированный в среду Protégé [30]; а также OntoSoft – система для описания и публикации научного программного обеспечения на основе онтологического подхода [31].

При выборе платформы для разработки онтологий важно учитывать её совместимость с существующими онтологиями, наличие поддержки импорта и экспорта данных в различных форматах и онтологических представлениях, а также наличие встроенных механизмов визуализации, интеграции и документирования. Особенно значимым является наличие механизмов повторного использования фрагментов онтологий и их семантической аннотации [32], что способствует устойчивости и расширяемости онтологических моделей.

Ниже представлен краткий обзор статей в области управления водными ресурсами и гидрологии. Проанализированные публикации касаются либо разработки онтологий, либо их использования. В таблице 1.2 представлены основные онтологии, связанные с водными ресурсами и гидрологией. Значительная часть данных онтологий разработана в рамках международных стандартов и инициатив, направленных на обеспечение совместимости геопространственных данных и формирование распределённых инфраструктур пространственной информации.

Одним из ключевых направлений развития семантических гидрологических моделей является обеспечение онтологического выравнивания (ontology alignment) и референсной интеграции (reference integration) между различными концептуальными схемами, что позволяет сопоставлять эквивалентные классы, свойства и отношения, используемые в различных стандартах и информационных системах. Реализация таких механизмов способствует повышению согласованности представления гидрологических знаний и обеспечивает возможность автоматизированного объединения гетерогенных наборов данных.

Использование семантических моделей позволяет поддерживать расширенные возможности интеллектуального поиска, интеграции и аналитической обработки данных в геоинформационных и веб-ориентированных системах.

Таблица 1.2 – Онтологии, связанные с водными ресурсами и гидрологией

Онтология	Контекст	Определение
HY_Features [33]	Международный стандарт, описывающий концептуальную модель поверхностных гидрологических объектов.	HY_Features – онтология поверхностных гидрологических объектов, описывающая ключевые сущности, их свойства и связи для моделирования и анализа водных систем в геоинформационных приложениях.
SWOP [34]	Surface Water Ontology Pattern – это онтологический шаблон, разработанный для представления знаний о поверхностных водах.	SWOP направлена на обеспечение семантической интероперабельности при моделировании поверхностных водных объектов, абстрактных путей потока и их пространственно-временных свойств.
SWEET [14]	Semantic Web for Earth and Environmental Terminology – это онтология для представления концепций, связанных с окружающей средой и гидрологией.	Это доменная онтология, охватывающая концепты поверхностных и подземных вод, предназначенная для поддержки описания, поиска и интеграции научных данных в области наук о Земле.
INSPIRE [35]	Infrastructure for Spatial Information in the European Community Ontology – разработана для создания стандартизированной инфраструктуры пространственных данных.	Представляет собой OWL-реализацию концептуальных моделей INSPIRE, охватывающую тематические области и обеспечивающую формальное описание понятий, отношений и ограничений, что облегчает автоматизированную интеграцию, обмен и интерпретацию пространственных данных между странами ЕС.
hydrOntology [36]	hydrOntology была предложена в 2009 году как средство интеграции гидрографической информации из различных источников данных и стандартов, таких как INSPIRE, ISO 19115 и WFD.	Это доменная онтология, предназначенная для описания гидрографических объектов (реки, озёра, бассейны и т. д.), обеспечивающая семантическое согласование и интероперабельность между различными гидрологическими системами.
Surface Water Ontology [37]	Surface Water Ontology была разработана для более точного описания поверхностных водных объектов и выравнивания понятий между различными онтологиями гидрографии.	Это онтология, основанная на Национальной гидрографической базе данных (NHD), направленная на формализацию типов и пространственных связей поверхностных водных объектов, таких как водотоки, водоёмы и водосборы.
HOW [38]	Hydrologic Ontology for the Web была представлена на конференции Американского геофизического союза в 2003 году с целью создания онтологической основы для публикации и совместного использования гидрологических данных	Представляет собой OWL-онтологию, описывающую основные концепции гидрологических данных, включая реки, притоки, бассейны, датчики и параметры измерений, и используется для стандартизации обмена информацией между веб-приложениями.

Продолжение Таблицы 1.2

HyFo [39]	HyFO была представлена в 2016 году как расширение фундаментальной онтологии DOLCE, с целью формализации понятий, связанных с подземными и поверхностными водами.	Это онтология, ориентированная на описание ключевых классов и связей между водными телами (подземными, поверхностными и атмосферными) в гидрологическом домене, служащая в качестве опорной модели для разработки стандартов, таких как GWML2.

Онтология HY_Features, представленная в международном стандарте OGC 14-111r6, служит концептуальной моделью для описания поверхностных гидрологических объектов. Её целью является обеспечение семантической совместимости и интеграции гидрологических данных, независимо от геометрии, источника или масштаба представления. Модель определяет основные сущности, такие как реки, озёра, водосборные бассейны, и описывает их атрибуты и взаимосвязи. В основе лежит абстрактный подход, при котором геометрия отделена от тематической информации, что упрощает повторное использование и интероперабельность данных. HY_Features активно используется в гидрологических моделях, платформах обмена данными и ГИС-приложениях. Онтология поддерживает привязку к другим стандартам, включая ISO 19156 (Observations and Measurements) и может использоваться в связке с RDF и OWL для семантической аннотации.

На рисунке 1.7 представлена структура модулей и пакетов модели HY_Features, иллюстрирующая организацию её компонентов и взаимосвязей.

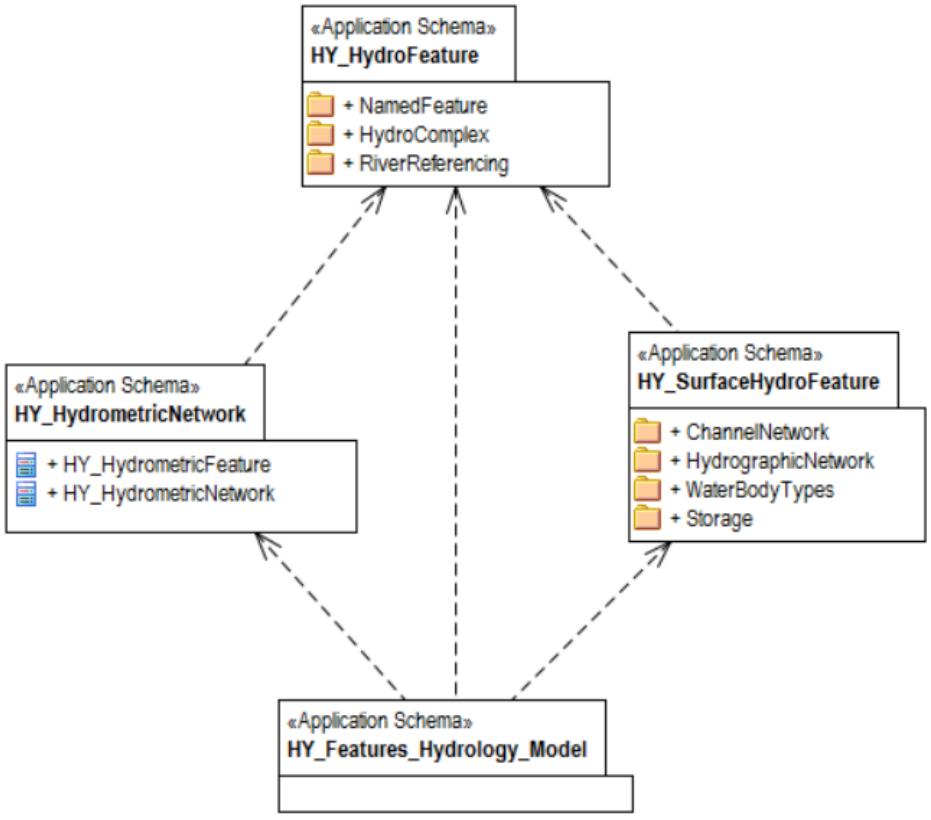


Рисунок 1.7 – Модули и пакеты HY_Features [33]

SWOP [34] разработан для извлечения знаний и служит концептуальной основой для более сложных онтологий в этой области. Он различает ландшафтные контейнеры (например, русла) и водоёмы (например, реки), которые их заполняют. Модули Dry и Wet описывают соответственно семантику сухих форм рельефа и водных тел. Шаблон реализован на языке OWL и предназначен для аннотирования данных поверхностных вод. Его практическая применимость показана через примеры семантических запросов к аннотированным геоданным.

Онтология SWEET [14] представляет собой онтологию среднего уровня для терминов в области окружающей среды, включающую гидрологическую предметную область.

INSPIRE Ontology [35] представляет собой OWL-представление концептуальных моделей, разработанных в рамках европейской инициативы Infrastructure for Spatial Information in the European Community (INSPIRE) для обеспечения совместимости пространственных данных. Она служит связующим звеном между геоинформационными стандартами и семантическими технологиями, позволяя интегрировать и использовать гидрологические и другие тематические данные в машиночитаемой форме. Онтология охватывает более 30 тематических доменов, включая «Гидрографию», «Зоны управления», «Экологическое состояние» и другие, и обеспечивает единообразие описания пространственных объектов, таких как реки, озёра и административные границы, в трансграничных приложениях. Благодаря использованию OWL и RDF форматов, INSPIRE Ontology способствует разработке интероперабельных систем, поддерживающих пространственные запросы, автоматическую аннотацию и интеграцию данных в рамках Европейского геопортала. Она активно применяется в проектах по мониторингу окружающей среды, управлению природными ресурсами и пространственному планированию. Таким образом, INSPIRE Ontology является ключевым инструментом для стандартизации и семантического согласования пространственных данных на европейском уровне. На Рисунке 1.8 представлена архитектура взаимодействия SPARQL-запросов с геопространственными метаданными, которая реализуется с использованием онтологии INSPIRE и промежуточного программного слоя (middleware).

hydrOntology создаёт основу для сравнения и согласования различных моделей данных, представленных в гидрологических информационных системах. Она включает в себя как терминологические элементы (словари и синонимы), так и иерархические связи между водными объектами. Основной задачей hydrOntology является поддержка геопространственной совместимости в рамках европейских и международных инициатив по управлению водными ресурсами. В работе Vilches-Blázquez et al. онтология также используется для оценки покрытия предметной области другими гидро-онтологиями, демонстрируя её роль как справочной модели.

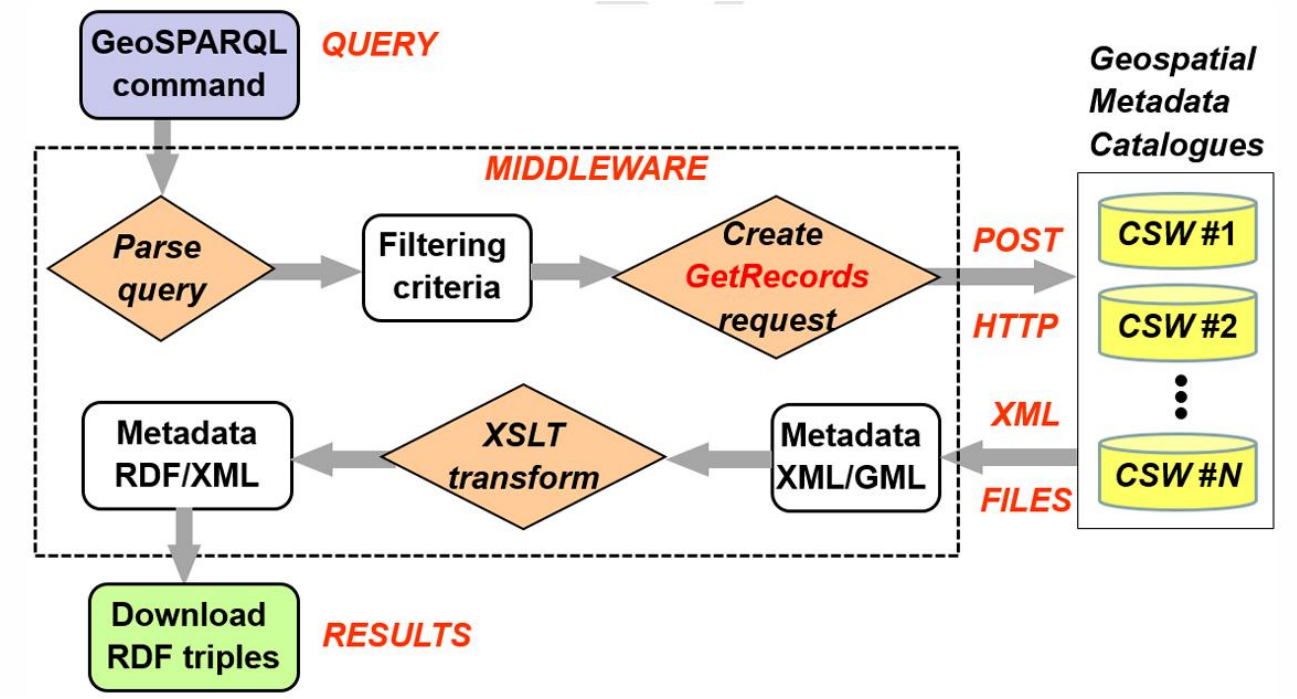


Рисунок 1.8 – Архитектура взаимодействия SPARQL-запросов с геопространственными метаданными [35]

Surface Water Ontology служит как инструмент согласования пространственных понятий при слиянии различных источников гидрографических данных. Особое внимание уделяется различению между водоносными объектами (реками, озёрами) и структурными элементами ландшафта (руслами, бассейнами). В Surface Water Ontology используются OWL-классы для определения семантики ландшафтных и водных компонентов, что позволяет автоматизировать семантические запросы и анализ геоданных. Она также может применяться для аннотирования геопространственных наборов данных в рамках более широких онтологических систем.

HOW Ontology предоставляет структуру, в которой гидрологические объекты и процессы могут быть описаны в машиночитаемом виде с помощью стандартов Semantic Web. Она способствует совместимости между гидрологическими веб-службами за счёт унифицированного словаря, что особенно важно при интеграции данных из различных источников. Онтология реализована с использованием OWL и поддерживает автоматизированный обмен и интерпретацию данных, связанных с мониторингом водных ресурсов, такими как потоки воды, осадки и уровни рек. HOW также демонстрирует потенциальную возможность связывать наблюдения в реальном времени с концептуальной моделью через Semantic Web технологии.

HyFO формализует различия между водоносными зонами, водоразделами, потоками и источниками воды, создавая основу для унифицированного моделирования гидрологических данных. Она расширяет существующие словари и стандарты, такие как GroundWater Markup Language (GWML2), добавляя семантически чёткие определения и иерархии. Эта онтология играет важную роль в обеспечении интероперабельности между различными системами

мониторинга и анализа водных ресурсов. В частности, HyFO помогает автоматизировать процессы оценки, симуляции и обмена данными между научными и управленческими структурами, работающими с водной средой.

С точки зрения охраны окружающей среды, онтологии в области гидрологии и управления водными ресурсами охватывают все смежные направления – климатологию, сельское хозяйство, водные науки – поскольку именно онтологии обеспечивают представление знаний, их совместное использование, управление данными и интеграцию знаний в интеллектуальные системы. Настоящее исследование представляет собой комплексный обзор онтологий, разработанных в области гидрологии.

Проведённый анализ показал, что в то время, как общее количество исследований, посвящённых разработке онтологий, стремительно растёт с каждым годом, области гидрологии и управления водными ресурсами не всегда успевают за этим прогрессом. Особенно тревожной является нестабильная динамика числа онтологических исследований в гидрологическом домене – несмотря на активное развитие киберинфраструктур, таких как информационные системы об окружающей среде, графы знаний и открытые онтологические сети (OKNs), ориентированных на решение экологических задач нового времени. Хотя в этих системах по-прежнему используются устаревшие онтологии, очевидно, что данная область требует дополнительных усилий и внимания со стороны исследовательского сообщества.

1.3 Методы семантического вывода и извлечения новых знаний в гидрологических онтологиях

Современные гидрологические онтологии служат не только средством унифицированного описания объектов и наблюдений, но и основой для семантического вывода, позволяющего выявлять неявные связи и получать новые знания из разнородных источников данных [40]. Концептуальная модель HY_Features [41] формализует сущности поверхностной гидрологии и их отношения, а стандарт SSN/SOSA [42] задаёт строгую схему описания сенсоров, наблюдений и наблюдаемых свойств, что создаёт прочный каркас для правил и онтологических ограничений. В совокупности эти спецификации поддерживают задачи от проверки согласованности гидрографической сети до обнаружения событий качества воды и трассировки потенциальных источников загрязнения на основе потоков наблюдений. Публикации последних лет показывают, что сочетание онтологических моделей с правилами вывода обеспечивает интеграцию данных, повышение интероперабельности и возможность автоматического получения новых фактов, критичных для мониторинга и управления водными ресурсами [43-44].

Язык правил Семантической паутины (SWRL) – это расширение стандарта OWL 2 DL, включающее подмножество RuleML, предназначенное для декларативного представления знаний с использованием правил в форме хорновских предложений [45]. Он объединяет возможности логики описаний (DL) и логики Хорна (HL), что позволяет формулировать сложные утверждения

о предметной области, выходящие за рамки простой декларации классов и свойств в OWL, а также дополняет онтологии механизмами семантического вывода [46]. SWRL обеспечивает более высокую выразительность, позволяя использовать имена классов как предикаты, формулировать равенства и неравенства, задавать конъюнкции атомов в антецеденте и консеквенте, переносить характеристики между классами и свойствами без создания иерархий подклассов, а также создавать новых индивидов с помощью экзистенциальных операторов [47-48]. Несмотря на это, классический SWRL не способен в полной мере моделировать сложные сценарии реального мира, такие как неопределённые знания, не-монотонные зависимости и удаление фактов, что обусловило разработку множества расширений его выразительности для повышения применимости в экспертных системах, интеграции гетерогенных данных и решении прикладных задач Семантической паутины [47].

Синтаксис SWRL базируется на модельно-теоретической семантике OWL и синтаксических конструкциях RuleML, а классические правила представляют собой положительные, не содержащие функциональных символов хорновские импликации, состоящие из предпосылки (тела правила) и следствия (головы правила) [45]. Онтология, поддерживающая SWRL, включает базу знаний OWL и набор аксиом в виде таких правил.

Аксиома-правило состоит из посылки (antecedent, тело) и заключения (consequent, голова), каждая из которых содержит (возможно пустое) множество атомов. Аксиоме-правилу также может быть присвоена ссылка-URI, которая может служить для идентификации правила.

rule ::= 'Implies(' [URIreference] { annotation } antecedent consequent ')'
antecedent ::= 'Antecedent(' { atom } ')'
consequent ::= 'Consequent(' { atom } ')'

В неформальной трактовке правило можно читать так: если посылка выполняется (является «истинной»), то должно выполняться и заключение. Пустая посылка считается тривиально выполняющейся (истинной), а пустое заключение – тривиально невыполняющимся (ложным). Правила с пустой посылкой могут использоваться для задания безусловных фактов; однако такие факты лучше выражать средствами самого OWL, то есть без конструкции правила.

Непустые посылки и заключения выполняются тогда и только тогда, когда выполняются все их составляющие атомы, то есть они трактуются как конъюнкции атомов. Как отмечалось выше, правила с конъюнктивными заключениями можно легко преобразовать (с помощью преобразований Ллойда–Топора [Lloyd87]) в несколько правил, каждое с атомарным заключением:

atom ::= description '(' i-object ')'
| dataRange '(' d-object ')'
| individualvaluedPropertyID '(' i-object i-object ')'
| datavaluedPropertyID '(' i-object d-object ')'
| sameAs '(' i-object i-object ')'
| differentFrom '(' i-object i-object ')'

| builtIn '(' builtinID { d-object } ')'

builtinID ::= URIreference

Атомы могут иметь вид:

- C(x) – где C является OWL-описанием или диапазоном данных;
- P(x, y) – где P является свойством OWL;
- sameAs(x, y) или differentFrom(x, y) – для задания равенства или различия объектов;
- builtIn(r, x, …) – где r – встроенное отношение.

В контексте OWL Lite описания в атомах вида C(x) могут быть ограничены только именами классов.

Неформальная интерпретация:

- Атом C(x) выполняется, если x является экземпляром класса или диапазона данных C.
- Атом P(x, y) выполняется, если x связан с y через свойство P.
- Атом sameAs(x, y) выполняется, если x и y интерпретируются как один и тот же объект.
- Атом differentFrom(x, y) выполняется, если x и y интерпретируются как разные объекты.
- Атом builtIn(r, x, …) выполняется, если встроенное отношение r (см. разделы 3 и 8) выполняется над интерпретациями аргументов.

Формы sameAs и differentFrom можно рассматривать как «синтаксический сахар»: они удобны, но не расширяют выразительную мощность языка (равенство/неравенство можно выразить комбинацией возможностей OWL и правил без явных атомов (не)равенства).

Для обеспечения разрешимости SWRL применяется ограничение DL-Safety [49], которое требует, чтобы все переменные в правилах появлялись в атомах, упоминающихся в антецеденте и относящихся к известным индивидуумам онтологии. Это предотвращает выводы о гипотетических объектах, не описанных явно в базе знаний, но ограничивает полноту выражения знаний.

В соответствии с OWL Semantics and Abstract Syntax, если дано отображение типов данных D, абстрактная OWL-интерпретация представляет собой кортеж:

I = <R, EC, ER, L, S, LV>

где:

- R – множество ресурсов;
- LV ⊆ R – множество значений-литералов;
- EC – отображение классов и типов данных в подмножества R и LV соответственно;
- ER – отображение свойств в бинарные отношения на R;
- L – отображение типизированных литералов в элементы LV;
- S – отображение имён индивидов в элементы EC(owl:Thing).

Для обработки встроенных отношений (built-ins) карта типов данных дополняется отображением встроенных отношений в кортежи над

соответствующими множествами. Например, op:numeric-add сопоставляется с тройками числовых значений, корректно интерпретирующими операцию сложения.

Разрешение изменения карты типов данных позволяет разным реализациям SWRL поддерживать разные встроенные отношения. Если реализация поддерживает конкретный тип данных, рекомендуется реализовать и встроенные операции для этого типа, указанные в разделе 8.

Дано абстрактное OWL-толкование I, *связка* B(I) – это абстрактная OWL-интерпретация, расширяющая I так, что S сопоставляет i-переменные с элементами EC(owl:Thing), а L – d-переменные с элементами LV.

Атом считается удовлетворённым интерпретацией I при выполнении условий из Таблицы условий интерпретации, где:

- C – описание класса в OWL DL,
- D – диапазон данных в OWL DL,
- P – свойство с индивидуальным значением,
- Q – свойство с значением-данными,
- f – встроенное отношение,
- x, y – переменные или индивиды OWL,
- z – переменная или значение данных OWL.

Семантические условия для аксиом и онтологий не изменяются:

- Интерпретация удовлетворяет онтологию тогда и только тогда, когда она удовлетворяет каждую аксиому (включая правила) и факт в онтологии;
- Онтология считается согласованной, если существует хотя бы одна интерпретация, которая её удовлетворяет;
- Онтология O2 является логическим следствием O1, если каждая интерпретация, удовлетворяющая O1, также удовлетворяет O2.

Простое использование правил – вывод свойства `hasUncle` из комбинации свойств `hasParent` и `hasBrother`:

Неофициально:

```
ruby
hasParent(?x1,?x2) ∧ hasBrother(?x2,?x3) ⇒
hasUncle(?x1,?x3)
```

В абстрактном синтаксисе:

```
scss
Implies(
  Antecedent(
    hasParent(I-variable(x1) I-variable(x2))
    hasBrother(I-variable(x2) I-variable(x3))
  )
  Consequent(
    hasUncle(I-variable(x1) I-variable(x3))
  )
)
```

Если Джон имеет Мэри как родителя, а Мэри имеет Билла как брата, то Джон имеет Билла как дядю.

SWRL обеспечивает не только декларативное представление фактов в онтологиях OWL, но и возможность их извлечения с помощью специализированного языка запросов SQWRL [50]. Среди ключевых преимуществ SWRL выделяют возможность:

a) использования имён классов и их описаний в качестве предикатов;

b) задания равенств и неравенств;

c) объединения атомов в конъюнкции как в теле, так и в голове правила;

d) переноса характеристик между классами и свойствами без введения дополнительных иерархий;

e) вывода новых индивидов с применением экзистенциальных операторов.

Эти свойства, дополняя выразительные возможности OWL, делают SWRL одним из ключевых инструментов построения онтологий с расширенным семантическим выводом и обеспечивают ему устойчивую поддержку в научном сообществе Семантической паутины. Основные термины и толкования SWRL:

- Clausal Logic (CL) представляет собой особый вид первого порядка логики (FOL) – мощного формализма, используемого для выражения связей между объектами в определенной области знаний с помощью квантованных переменных и предикатов. CL служит основой для логического программирования (LP) через свой фрагмент, называемый логикой кодов Хорна (HCL).
- Глава Хорна – это правило логической программы, содержащее дизъюнкции литералов (атрибутов правил) с не более чем одним положительным литералом [51]. Конечное множество глав Хорна вместе с абсолютными фактами называется логической программой (LP). Проще говоря, LP – это конечный набор правил и фактов.
- Datalog, с другой стороны, представляет собойHorn-прологовые правила без функций, содержащие только переменные и константы в качестве терминов.
- Description Logic (DL) [52] – по сути, допустимый фрагмент FOL, который позволяет моделировать области знаний с помощью концептов (классов), бинарных отношений между ними (называемых ролями или свойствами) и индивидуальных экземпляров (фактов). Будучи разрешимым, DL служит основой для большинства языков моделирования доменов (онтологий), включая OWL.
- Decidability (разрешимость): в контексте данной работы разрешимость означает способность «Интерпретатора» (Reasoner) осуществлять вывод – проверять согласованность логических следствий аксиом онтологии и возвращать значение true или false в ограниченное время. То есть, для любой онтологии, набора правил и предложения Reasoner способен проверить, является ли предложение логическим следствием онтологии и правил. Более того,

существует останавливающаяся процедура.

- OWL-DL представляет собой разрешимый (decidable) фрагмент языка онтологий Web Ontology Language (OWL). OWL – это наиболее выразительный язык онтологий для Семантической Web, и по рекомендации W3C [45] он развивается со временем, совершенствуя такие языки, как RDF, FRAMES, DAML и OIL. Другие фрагменты OWL включают OWL-Full, который превосходит OWL-DL по степени выразительности, и OWL-Lite, который является наименее выразительным. Для осуществления вывода в онтологиях OWL извлекается подкласс OWL-Full, полностью соответствующий рамкам DL, – это и есть OWL-DL. Обычно онтологии OWL ограничиваются степенью выразительности DL, чтобы обеспечить разрешимость, а полноценная OWL-DL онтология может иметь выразительность, эквивалентную уровню SHOIN(D) по метрике DL.
- Метрика выразительности DL: В языках на основе DL фактические знания, такие как определения терминов и индивидуальные утверждения, обычно хранятся в виде формул первого порядка (FOL). Однако к этим формулам накладываются ограничения, чтобы обеспечить разрешимость и эффективность вывода по онтологии [43]. Эти ограничения также задают уровень выразительности языка DL по сравнению с FOL, которая является очень выразительной, однако недопустимой для автоматического вывода (недекдурируемой).

В работе [53] представлен онтологический фреймворк, в котором интеграция OWL и SWRL используется для моделирования, хранения и автоматической обработки знаний, связанных с управлением качеством воды. Предложенный подход позволяет не только описывать текущие состояния и характеристики водных объектов, но и формулировать условия, при которых система автоматически выявляет нарушения, прогнозирует развитие ситуации и предлагает возможные меры реагирования.

Разработка онтологической модели для управления качеством водных ресурсов предполагает комплексное отражение как структурных элементов предметной области, так и логико-вычислительных возможностей системы. Основными требованиями к такой онтологии являются:

1. Интеграция разнородных источников данных – онтология должна обеспечивать единое представление информации, поступающей из гетерогенных источников: датчиков и сенсорных узлов, автоматических пробоотборников, лабораторных отчётов, нормативно-правовых документов, а также экспертных заключений.
2. Семантическая репрезентация наблюдений и устройств – модель должна описывать результаты наблюдений (параметры качества воды, время и место измерений, тип используемого оборудования), а также сведения об устройствах (тип, местоположение, идентификаторы, например RFID).
3. Нормативно-ориентированная классификация – онтология должна обеспечивать возможность автоматической классификации состояния водных

объектов в соответствии с несколькими нормативными системами (международные, региональные и локальные стандарты качества воды).

4. Поддержка механизмов логического вывода – модель должна интегрировать правила семантического вывода (например, SWRL) для автоматического выявления ситуаций, требующих внимания, включая превышение допустимых концентраций, установление вероятных источников загрязнения и генерацию предупреждений.

5. Расширяемость и модульность – архитектура онтологии должна предусматривать добавление новых параметров, правил, классов и взаимосвязей без необходимости кардинальной перестройки модели.

В соответствии с положениями [54], знания, связанные с мониторингом качества воды, включают три ключевых компонента: Наблюдательные данные, формируемые сенсорными устройствами и автоматическими пробоотборниками (например, значения концентрации аммиака, pH, температуры и других показателей качества воды), нормативно-правовые сведения, содержащие установленные регулирующими органами предельно допустимые значения параметров воды (например, стандарты безопасного питьевого водоснабжения), знания о водной среде, аккумулируемые в результате исследований (в том числе сведения о загрязнителях, их типах и характеристиках водных объектов).

1. В данной работе предложено расширение базовой модели за счёт включения информации об источниках загрязнения, что позволило выделить четыре взаимосвязанных модуля онтологии:

2. Базовая онтология [55] – включает классы и отношения, необходимые для интеграции данных о качестве воды в реальном времени, поступающих из различных источников (сенсоров или лабораторных измерений).

3. Онтология нормативов [56] – модуль, описывающий допустимые пороговые значения параметров воды в соответствии с требованиями различных регулирующих органов.

4. Онтология источников загрязнения – модуль, представляющий загрязнителей как сущности и описывающий их атрибуты.

5. Онтология экспертных правил – модуль, содержащий правила типа *если–тогда* для экспертного анализа данных о воде.

6. Интеграция указанных модулей в единую онтологическую структуру обеспечивает возможность комплексного семантического вывода.

Так, данные наблюдений, поступающие в базовый модуль, сопоставляются с онтологией нормативов посредством экспертных правил, что позволяет автоматически классифицировать состояние водных объектов по установленным стандартам. Дополнительно, правила, основанные на онтологии источников загрязнения, используются для идентификации вероятных причин обнаруженных нарушений. Такой подход обеспечивает не только фиксацию факта превышения пороговых значений, но и формирование причинно-следственных связей, что существенно повышает аналитическую ценность системы.

Базовая онтология фиксирует такие характеристики, как тип и единица измерения показателя, пространственно-временные атрибуты (координаты сенсорных узлов, момент проведения наблюдения), а также, при необходимости, сведения о применённых методах тестирования и используемом оборудовании.

В качестве верхнеуровневой онтологии выбрана SSN (Semantic Sensor Network Ontology), поскольку она позволяет формализовать описание датчиков в терминах их функциональных возможностей, процессов измерения, наблюдений и развертываний. Этот выбор обусловлен её высокой адаптивностью для моделирования беспроводных сенсорных сетей (WSN) и совместимостью с существующими стандартами представления данных. Для повышения полноты представления домена модель была расширена дополнительными классами и свойствами в соответствии с системными требованиями. Так, для описания временных характеристик интегрирована онтология OWL Time [57], а для задания географических координат – базовый словарь геолокаций [58].

Поскольку в рамках системы предполагается использование правил SWRL [45], разработка онтологии осуществлялась в среде *Protégé* версии 3.5, поддерживающей встроенные функции SWRL. Ограничения прямого импорта OWL2-версии SSN в Protégé 3 были обойдены путём выборочного включения необходимых концептов SSN с их последующим расширением за счёт других онтологий и собственных классов.

Для представления системных оповещений повторно использованы классы *event:Alert* из онтологии CSIRO и его специализированные подклассы (*event:EmailAlert*, *event:SMSAlert*). Классификация оборудования реализована через введение класса *DeviceType* и его подклассов (например, *AutoSampler* для автоматических пробоотборников). Связь объектов с конкретными устройствами отражается свойством *hasDevice*.

Пространственное расположение компонентов WSN формализуется с помощью классов *SensingNode*, *GatewayNode* и *CentralMonitoringNode* – подклассов ssn:Platform. Для случаев, когда сенсорный узел имеет несколько географических точек, введён класс *SensingNodeLocation*, являющийся подклассом twcc:MeasurementSite и, в свою очередь, подклассом geo:Point. В соответствии с определёнными требованиями, каждый экземпляр *SensingNode* обязан иметь местоположение, представленное экземпляром *SensingNodeLocation*, включать как минимум один сенсорный зонд и один RFID, но не более одного автоматического пробоотборника.

Подкласс *WaterQuality* (наследуемый от ssn:Property) расширен классом *RiversWaterQuality* для категоризации элементов качества воды на биологические, гидроморфологические и физико-химические. В рамках класса ssn:Sensor выделены специализированные подклассы, отражающие измерение конкретных показателей (например, *DissolvedOxygenSensor* для датчиков растворённого кислорода). Прибор, фиксирующий несколько параметров, может быть отнесён сразу к нескольким таким подклассам.

В ssn:Observation добавлено объектное свойство *observationResultLocation*

для фиксации места проведения наблюдения, сопровождаемое аксиомами *observationResultLocation only geo:Point* и *observationResultLocation min 0*. В классе ssn:Platform установлено ограничение *dul:attachedSystem owl:hasValue InWaterSense*, указывающее на принадлежность всех экземпляров платформ системе InWaterSense.

Дополнительно введён подкласс ssn:FeatureOfInterest – *WaterFeature*, который в первом кейсе будет содержать экземпляр *RiversWaterFeature*, а во втором – *DrinkingWaterFeature*. Такое расширение обеспечивает полноту представления пространственных, временных и функциональных характеристик сенсорных узлов, что является необходимым условием для реализации механизмов семантического вывода в задачах мониторинга качества воды.

Онтология нормативов база WFD (Water Framework Directive) предусматривает классификацию показателей качества воды по трем основным категориям: биологические, гидроморфологические и физико-химические [59]. Вместо бинарной оценки («чистая/загрязненная»), как это предлагалось в [54], применяется пятиуровневая шкала экологического состояния водных объектов с привязкой к цветовой маркировке:

- высокий статус (синий);
- хороший (зеленый);
- умеренный (желтый);
- плохой (оранжевый);
- очень плохой (красный).

При этом используется принцип *one-out-all-out*, согласно которому общий экологический статус определяется по наименее благоприятному элементу качества водного объекта [59].

Для формального описания всех пяти категорий введен класс *WFDSurfaceWaterStatus*, содержащий отдельные подклассы для каждого статуса. Соответствие между статусом и цветовым кодом фиксируется посредством свойства owl:equivalentClass (например, High owl:equivalentClass Blue).

Чтобы отразить применимость статусов WFD к конкретным категориям качества воды (*RiversWaterQuality*), разработан дополнительный класс *EcologicalStatus*. Ограничение owl:Restriction задает, что свойство *hasStandard* должно ссылаться исключительно на значения, относящиеся к классу WFD.

Класс twcc:WaterMeasurement используется в качестве обобщенного для всех элементов, описывающих статус качества воды, таких как *HighNutrientConditions*. Связь между данными наблюдений и статусами измерений задается через свойство *twcc:hasMeasurement* из онтологии нормативов TWC-SWQP.

В работе [54] статусы фиксировались на уровне OWL-ограничений свойств. Однако в паттерне проектирования SSN такая реализация невозможна, поскольку отдельный поток данных от сенсора представлен несколькими связанными индивидами (например, ssn:Observation, geo:Point, time:Instant и др.).

Для решения этой задачи целесообразно использование правил SWRL, поддерживающих свободные переменные. Так, правило WFD:

«Если среднее значение общего аммиака менее 0,04, то водный объект относится к высокому статусу по условиям содержания питательных веществ» может быть представлено следующим образом:

ssn:Observation(?x) ∧ ssn:observedProperty(?x, Ammonia) ∧
ssn:observationResultTime(?x, ?y) ∧ hasObservationTime(?y, ?z) ∧
temporal:after(?z, "2013-02-13T09:11:00") ∧
ssn:observationResult(?x, ?r) ∧ ssn:hasValue(?r, ?v) ∧
dul:hasDataValue(?v, ?val) ∧ sqwrl:makeSet(?sv, ?val) ∧
sqwrl:avg(?avg, ?sv) ∧ swrlb:greaterThan(?avg, 0.04) →
HighNutrientConditions(?o)

Данное правило реализует последовательную проверку наблюдений, фильтрацию по дате, агрегацию и расчет средних значений, отбор данных, превышающих пороговое значение, и их классификацию в соответствующий класс.

В результате анализа существующих подходов установлено, что методы семантического вывода и извлечения новых знаний в гидрологических онтологиях обеспечивают более глубокое понимание и интерпретацию данных мониторинга водных ресурсов. Применение правил, формализованных в рамках стандартов Семантического Веба, таких как SWRL, позволяет автоматизировать выявление закономерностей и аномалий, что повышает оперативность принятия управленческих решений. Использование модульной структуры онтологий, включающей ядро наблюдений, нормативные требования, описание источников загрязнения и экспертные правила, обеспечивает комплексное представление предметной области. Семантический вывод на основе интегрированных онтологий позволяет классифицировать водные объекты в соответствии с международными и национальными стандартами качества. Дополнительно, моделирование причинно-следственных связей между параметрами качества воды и потенциальными источниками загрязнения способствует выявлению скрытых факторов деградации экосистем. Применение онтологий в сочетании с сенсорными сетями расширяет возможности автоматизированного анализа, позволяя переходить от сырых данных к обоснованным выводам. Такой подход способствует построению интеллектуальных систем мониторинга, способных адаптироваться к изменяющимся условиям и обновлять знания в режиме реального времени. Внедрение семантических методов также открывает перспективы интеграции разнородных источников данных и их унификации в едином формате. В конечном счёте, развитие и применение данных методов повышает точность прогнозов, надежность оценки состояния водных объектов и эффективность природоохранных мероприятий. Таким образом, семантический вывод и извлечение новых знаний в гидрологических онтологиях являются ключевыми компонентами современной экологической аналитики и цифрового управления водными ресурсами.

2 СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ТАБЛИЧНЫХ ДАННЫХ ДЛЯ ПОСТРОЕНИЯ И ПОПОЛНЕНИЯ ОНТОЛОГИЙ В МОНИТОРИНГЕ ВОДНЫХ РЕСУРСОВ

Извлечение осмысленных знаний из таблиц с учётом онтологий или графов знаний определяется как Semantic Table Interpretation (STI) или семантическая аннотация таблиц [60]. Цель данного обзора - предоставить всесторонний и современный анализ методов и задач, применяемых для выполнения STI. В данном разделе приведены пять основных подзадач STI, хотя существующие исследования в основном сосредоточены лишь на трёх из них.

Рассмотренные подходы систематизированы в три крупные группы:

1) эвристические методы;

2) методы на основе признаков (feature engineering);

3) методы глубокого обучения [62].

В обзоре анализируются их преимущества и ограничения на основе существующих наборов данных и бенчмарков. Завершающая часть статьи выделяет нерешённые научные задачи и барьеры, которые необходимо преодолеть для достижения полностью автоматической интерпретации любых типов таблиц, встречающихся в сети Интернет.

Основная идея семантической интерпретации таблиц заключается в установлении соответствия между элементами таблицы и сущностями, понятиями или связями, описанными в графах знаний (Knowledge Graphs, KG). Эти графы могут быть как универсальными (например, DBpedia [63] и Wikidata [64]), так и специализированными для конкретных организаций.

Графы знаний (KGs) являются основой семантической интерпретации, так как описывают объекты реального мира и их взаимосвязи [65, 66]. Каждый элемент графа представлен тройкой *⟨субъект, предикат, объект⟩*. Например: ⟨Paris, hasPopulation, 2m⟩, где предикат hasPopulation связывает сущность Paris со значением 2 миллиона. Различают несколько типов KGs:

- Отраслевые (domain-specific) – описывают конкретную область знаний, например, сельское хозяйство [67], биотехнологии [68], экономику [69]. Они обеспечивают высокую точность, но ограниченное обобщение.
- Энциклопедические (encyclopedic) – крупные, многодоменные графы, такие как DBpedia [63], Wikidata [64], Freebase [70], YAGO [71].
- Бытовые или общезначимые (common-sense) – отражают общие знания о мире.

DBpedia – один из центральных узлов Linked Open Data Cloud [141], содержащий миллионы сущностей и сотни миллионов фактов, извлечённых из Wikipedia. Wikidata – крупнейший открытый проект с более чем 90 миллионами сущностей и 350 языками, обеспечивающий многоязычную поддержку и уникальные идентификаторы. YAGO (Max Planck Institute) объединяет данные Wikidata и WordNet, обеспечивая логическую целостность и высокую связность. Freebase, ранее созданный Metaweb и приобретённый Google, прекратил работу в 2015 году, а его данные были интегрированы в Wikidata.

Графы знаний применяются для сопоставления элементов таблицы с

сущностями KGs, что позволяет выявлять скрытые связи и уточнять семантику данных. Выбор подходящего KG зависит от степени пересечения его содержания с данными таблицы: чем больше пересечение, тем проще выполнить точное сопоставление. На практике оптимальный результат достигается при комбинировании нескольких графов знаний, что повышает как покрытие, так и детализацию интерпретации.

Семантическая аннотация может применяться для улучшения поисковых систем по наборам данных [73,74], совершенствования вопросно-ответных систем [75-77], обогащения баз знаний [78-80] и рекомендации наборов данных [81]. Появление специализированных поисковиков, таких как Google Dataset Search [82,83], является ярким примером практического внедрения подобных подходов.

Несмотря на прогресс, машинная интерпретация табличных данных остаётся нетривиальной задачей: существующие работы обычно охватывают лишь отдельные аспекты – извлечение и обогащение веб-таблиц [84], преобразование и понимание таблиц [85], применение глубокого обучения к табличным данным [86,87], информационное извлечение в сети [88] – без целостного рассмотрения полного конвейера семантической интерпретации. Анализ фокусировался на публикациях 2000–2021 гг.; поисковые запросы формировались из двух групп терминов: (i) по задачам (“semantic annotation”, “semantic interpretation”, “entity linking”, “type prediction”, “relation extraction”, “knowledge graph mapping” и др.) и (ii) по области (“table”, “web table”, “tabular data”, “structured data”), например: “web table semantic annotation” (также учитывались материалы SemTab [89]). Для каждой комбинации отбирались топ-10 цитируемых работ, их аннотации просеивались по критерию наличия методов сопоставления элементов таблицы с графом знаний или описания новых датасетов; затем корпус расширялся по сетям со-цитирования тем же способом. В итоге рассмотрено 48 поисковых комбинаций; отобрано 38 статей на первом этапе и ещё 58 – на втором, дополнительно включены 7 работ с arXiv и 22 публикации SemTab; всего 114 релевантных источников из >700 просмотренных. Распределение по годам (рис. 2.1) демонстрирует заметный рост интереса к теме STI за последнее десятилетие.

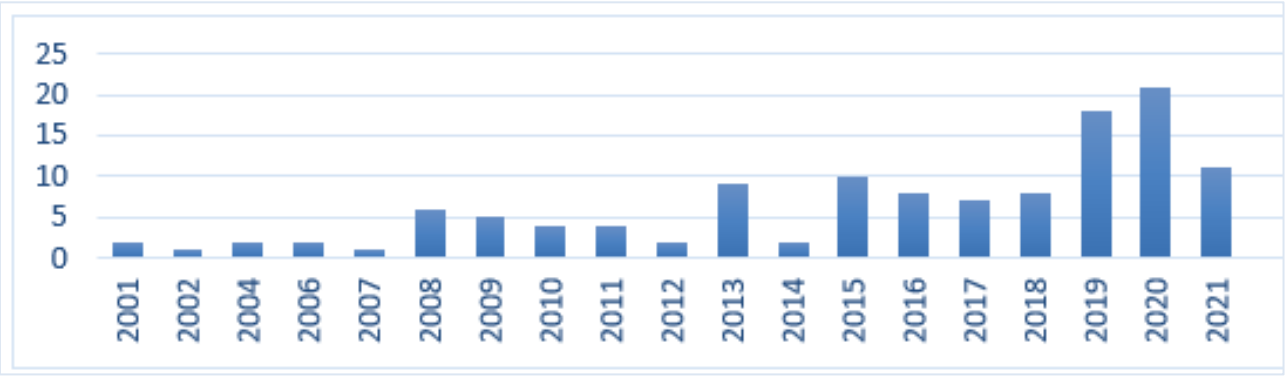


Рисунок 2.1 – Распределение статей по годам публикации

Каждая статья была вручную помечена по шести категориям (см. Рис 2.2), при этом некоторые работы охватывают сразу несколько направлений. Так, в [89] представлено и описание набора данных, и метод его аннотации. Хотя основной акцент данного обзора сделан на системах STI, он также затрагивает смежные вопросы – необходимость графов знаний и схем для привязки табличных данных к семантическим структурам, а также предобработку таблиц с применением методов NLP, повышающую точность интерпретации.

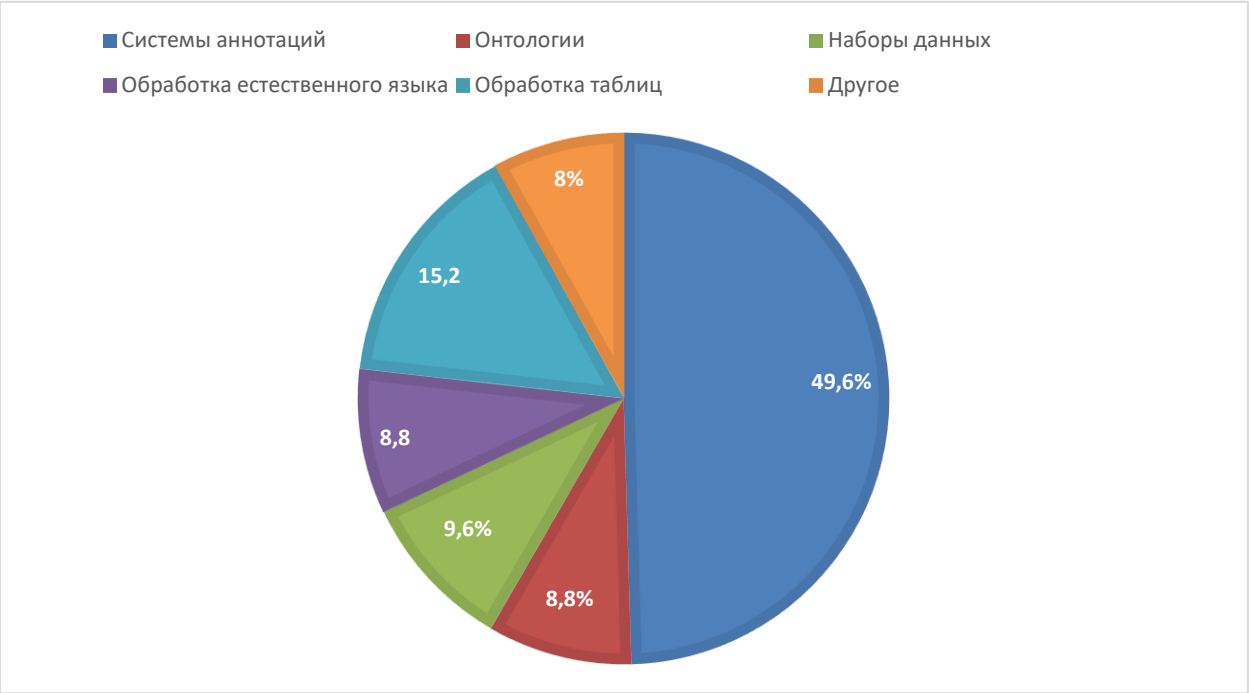


Рисунок 2.2 – Распределение тем статей

2.1 Типология и формальная модель таблиц в задаче семантической интерпретации

Поскольку таблицы различаются по структуре, источнику происхождения и назначению, авторы предлагают новую детализированную классификацию, основанную на существующих подходах, но с углублённым анализом реляционных таблиц. Помимо самих данных, важно учитывать метаданные и контекст, в котором таблица используется. Например, если таблица размещена на веб-странице, посвящённой Бундеслиге, то её содержимое, вероятнее всего, относится к футболу. Следовательно, при извлечении данных важно сохранять не только структуру таблицы, но и связанные с ней метаданные. STI использует графы знаний (Knowledge Graphs, KG) как источник справочной информации и основу для аннотирования.

Таблица представляет собой двумерную структуру данных, состоящую из n строк и m столбцов, что обеспечивает компактную и наглядную форму представления информации. Базовой единицей таблицы является ячейка, обозначаемая как Tij (где $0 \leq i < n$, $0 \leq j < m$), то есть элемент, расположенный на пересечении i-й строки и j-го столбца таблицы T. На Рис.2.3 представлена типология таблиц, используемая при семантической интерпретации. Таблицы

делятся на две категории:

1. Таблицы макета используются для визуального оформления веб-страниц. Они не содержат логически связанных данных и служат исключительно для улучшения восприятия интерфейса. На рисунке представлена такая таблица из интерфейса Amazon (справа).
2. Подлинные (истинные) таблицы содержат осмысленные данные, представленные в виде строк и столбцов. Они обладают синтаксической и семантической согласованностью, что делает их пригодными для автоматического извлечения знаний (STI). Примеры: таблица с ценой и таблица характеристик товара (в центре и слева на изображении).

В отличие от подлинных, таблицы макета не подходят для построения графа знаний, так как не содержат структурированных знаний.



Рисунок 2.3 – Классификация типов таблиц с более детализированным анализом подлинных таблиц по трём измерениям: структуре, внутренним взаимосвязям и ориентации.

Например, на Рис. 2.4 показаны подлинные таблицы, в которых указаны такие атрибуты продукта, как цена и цвет. Подлинные таблицы содержат реляционные данные, пригодные для машинной интерпретации и, следовательно, могут использоваться в процессе STI.

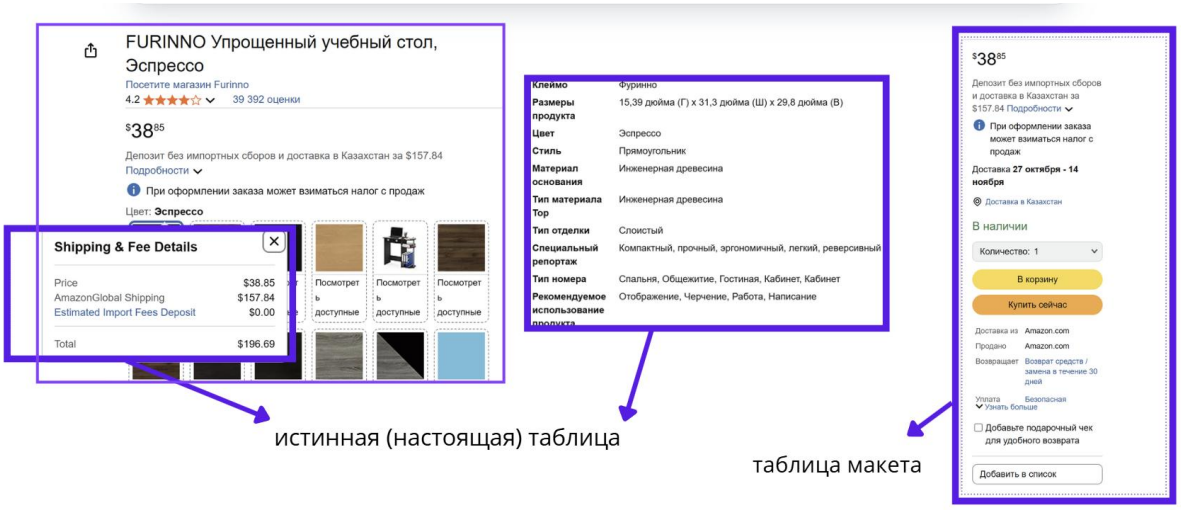


Рисунок 2.4 – Пример подлинных и макетных таблиц

Поэтому в данной работе предлагается классификация подлинных таблиц по трём невзаимоисключающим измерениям: структура**,** внутренние отношения и ориентация**.** Тип таблицы формируется как комбинация этих трёх характеристик.

В работе [90] основное внимание уделяется структуре макета таблицы, которая отражается прежде всего в составе её элементов. В соответствии с этим, подкласс «Измерение структуры» в предлагаемой классификации делится на четыре типа таблиц, представленных на рисунке 2.4:

1. Вложенные таблицы (Nested tables) содержат одну или несколько таблиц, размещённых внутри ячеек основной таблицы (Рис.2.5(a)).
2. Разделённые таблицы (Split tables) состоят из нескольких подтаблиц, ячейки которых не связаны между собой. Например, на рисунке 2.5(b) инфобокс города Чикаго состоит из нескольких подтаблиц: каждая из них описывает отдельный аспект основного объекта – местоположение, площадь и численность населения.
3. Компактные таблицы (Concise tables) содержат объединённые ячейки, что позволяет избежать повторений одинаковых значений в строках или столбцах. В таблице, представленной на рисунке 2.5(c), первая ячейка столбца lines объединяет шесть ячеек с одинаковым значением BART main lines.
4. Мультизначные таблицы (Multivalued tables) включают несколько значений в одной ячейке. К примеру, на рисунке 2.5(d) столбец lines used содержит в ячейках список маршрутов линий, что демонстрирует множественность значений в пределах одной ячейки.

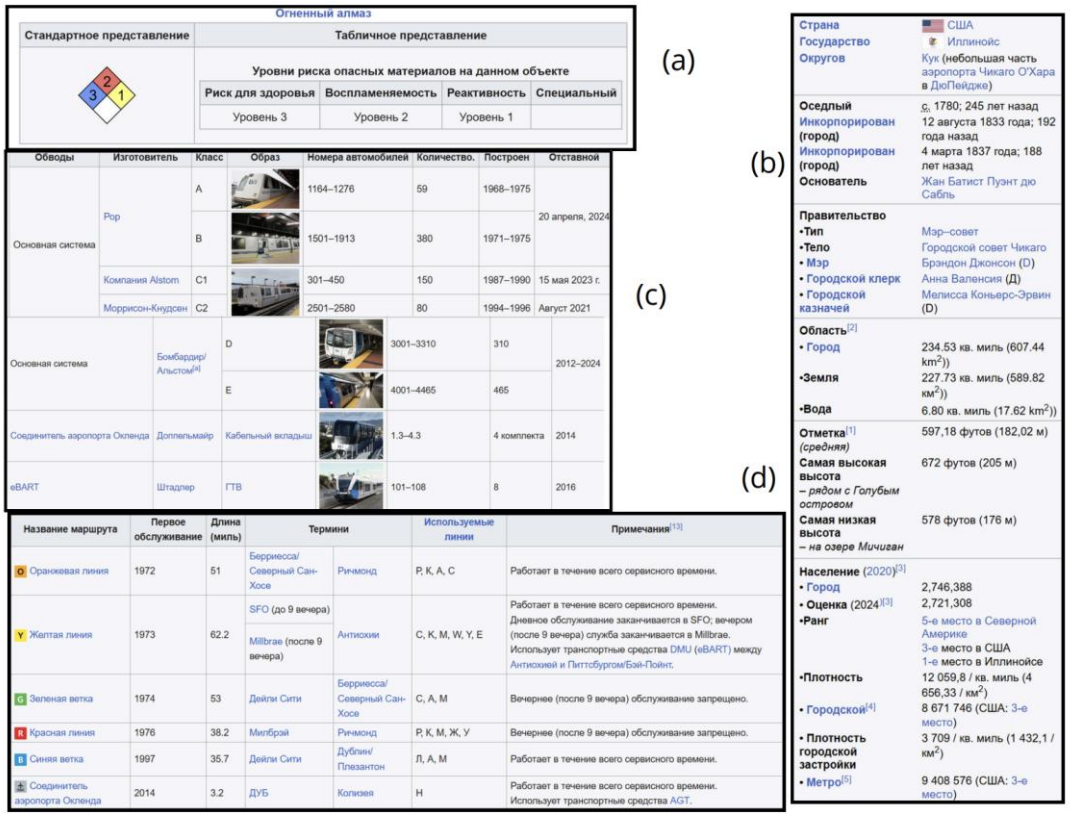


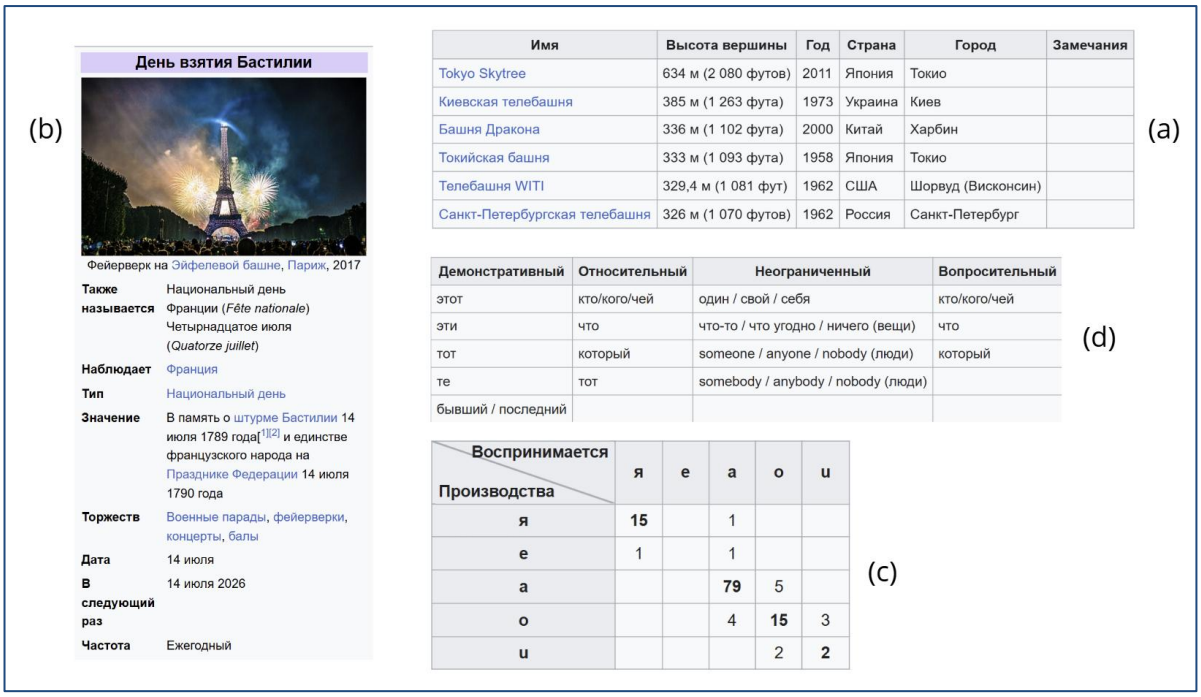
Рисунок 2.5 – Примеры различных структур таблиц:(a) вложенная таблица, (b) разделённая таблица, (c) компактная таблица (с объединёнными столбцами), (d) многозначная таблица (с повторяющимися строками).

Измерение внутренних взаимосвязей позволяет определить топологию семантических связей между ячейками таблицы. Одной из первых детальных классификаций таблиц с реляционными знаниями является работа [91], где выделены списки, атрибутно-значенческие таблицы, матрицы, перечни и формы. В [92] дополнительно выделяются календарные и перечислительные формы. Авторы [93, 94] расширили классификацию, введя понятие таблиц сущностей, классифицируя списки как реляционные структуры.

На основе этих подходов предлагается классификация таблиц:

1. Реляционные таблицы (Relational tables) представляют сведения об объектах и их атрибутах в строках и столбцах. При этом различают:
   i. Горизонтальные таблицы – сущности представлены по строкам, атрибуты – по столбцам.
   ii. Вертикальные таблицы – напротив, атрибуты по строкам, а сущности – по столбцам. Пример: на рисунке 6(a) показана горизонтальная таблица с описанием башен (высота, год постройки, страна, город).
2. Таблицы сущностей (Entity tables) описывают одну конкретную сущность через пары атрибут–значение. Пример: инфобокс с характеристиками "Bastille Day" на рисунке 6(b).
3. Матрицы (Matrix tables) представляют данные, зависящие от комбинации строки и столбца, часто используются для числовых или логических значений. Пример: матрица восприятия гласных на рисунке 6(c).

4. Прочие осмысленные таблицы (например, календари и перечисления) также могут нести семантическую нагрузку. Пример: перечисление местоимений на рисунке 6(d).



(b)

День взятия Бастилии

Фейерверк на Эйфелевой башне, Париж, 2017

Также называется	Национальный день Франции (*Fête nationale*) Четырнадцатое июля (*Quatorze juillet*)
Наблюдает	Франция
Тип	Национальный день
Значение	В память о штурме Бастилии 14 июля 1789 года[1][2] и единстве французского народа на Празднике Федерации 14 июля 1790 года
Торжеств	Военные парады, фейерверки, концерты, балы
Дата	14 июля
В следующий раз	14 июля 2026
Частота	Ежегодный

(a)

Имя	Высота вершины	Год	Страна	Город	Замечания
Tokyo Skytree	634 м (2 080 футов)	2011	Япония	Токио	
Киевская телебашня	385 м (1 263 фута)	1973	Украина	Киев	
Башня Дракона	336 м (1 102 фута)	2000	Китай	Харбин	
Токийская башня	333 м (1 093 фута)	1958	Япония	Токио	
Телебашня WITI	329,4 м (1 081 фут)	1962	США	Шорвуд (Висконсин)	
Санкт-Петербургская телебашня	326 м (1 070 футов)	1962	Россия	Санкт-Петербург	

(d)

Демонстративный	Относительный	Неограниченный	Вопросительный
этот	кто/кого/чей	один / свой / себя	кто/кого/чей
эти	что	что-то / что угодно / ничего (вещи)	что
тот	который	someone / anyone / nobody (люди)	который
те	тот	somebody / anybody / nobody (люди)	
бывший / последний			

(c)

Производства \ Воспринимается	я	е	а	о	u
я	**15**		1		
е	1		1		
а			**79**	5	
о			4	**15**	3
u				2	**2**

Рисунок 2.6 – Примеры различных типов внутренних связей в таблицах:(a) горизонтальная реляционная таблица, (b) таблица сущностей, (c) матричная таблица, (d) таблица перечислений, относящаяся к категории других подлинных таблиц

В ряде исследований реляционные таблицы рассматриваются как высший уровень в иерархии таблиц [93-95], однако их структура может варьироваться, особенно по способу представления субъектов.

В нашем исследовании особое внимание уделяется проблеме потери контекста при аннотировании таблиц, на которую неоднократно указывали предыдущие работы в области STI [84]. Мы исходим из того, что таблица сама по себе не является изолированным источником знаний: её смысл и назначение раскрываются через контекст окружения, в котором она размещена.

Под метаданными таблицы мы понимаем совокупность вспомогательных данных, извлекаемых из внешних или связанных источников, которые обеспечивают более полное понимание содержания таблицы и её структуры. Такие данные могут описывать характеристики, происхождение, логику построения и контекст использования таблицы, что позволяет эффективнее организовывать, извлекать, сохранять и управлять знаниями.

Задача аннотирования определяется элементами таблицы, которые необходимо аннотировать, а также типом кандидатов (индивиды, концепты или свойства из KG). В работе предлагается разделить процесс семантической интерпретации таблиц (STI) на пять основных задач: аннотирование ячеек сущностями, аннотирование типа столбцов, аннотирование свойств столбцов

[96], тематическое аннотирование и сопоставление строк с объектами [97] (иллюстрировано на Рис. 2.7).

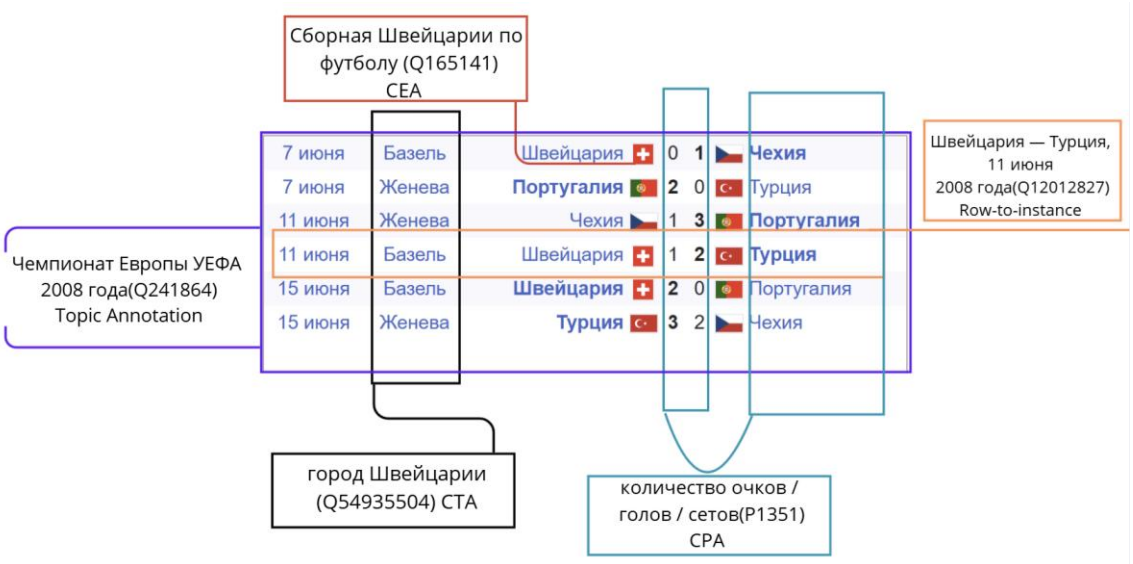


Рисунок 2.7 – Иллюстрация пяти задач STI для таблицы, описывающей результаты группы А чемпионата Европы УЕФА 2008 года

Аннотирование таблиц в контексте семантических графов сводится к согласованному решению взаимосвязанных задач. Аннотирование ячеек сущностями (CEA) трактуется как связывание текстовых упоминаний с объектами графа знаний (KG), тогда как аннотирование типа столбцов (CTA) требует выбора репрезентативного уровня обобщения в иерархии типов: чрезмерно общий тип снижает информативность, а слишком узкий – ухудшает покрытие значений столбца. Аннотирование свойств столбцов (CPA) определяет предикат для пары столбцов, согласующийся с онтологией целевого KG; тематическое аннотирование соотносит всю таблицу с одной сущностью/концептом; задача Row-to-Instance интерпретирует строку как описание одной сущности и позволяет выявлять «скрытый субъект» строки, недоступный при изолированном CEA. Типовая архитектура STI реализует многоэтапный конвейер: предобработка, генерация кандидатов, контекстно-зависимая обработка элементов таблицы и итеративная дезамбигуация, причем порядок шагов и возвраты к предыдущим стадиям варьируются [78, 98]. При вычислении достоверности аннотаций используются согласованные сигналы шести источников контекста: самой целевой ячейки (строковое/лексическое сходство) [99, 100], строки (согласованность атрибутов сущности), столбца (однородность типа значений), межстолбцовой структуры (глобальные зависимости таблицы), заголовков (семантические подсказки о типах/свойствах) и внешнего окружения таблицы (метаданные, гиперссылки, околотекст; включая межтабличные связи) [101, 102]. Интеграция этих сигналов позволяет строить стратегии интерпретации на уровнях строки, столбца и всей таблицы, в том числе с использованием моделей глубокого обучения для контекстуализированных представлений [101, 103]. Ключевым механизмом повышения качества служит

итеративная дезамбигуация: результаты одной задачи (например, CTA, указывающая, что столбец описывает «города») ограничивают пространство кандидатов другой (CEA, исключая не-города), после чего цикл CEA/CPA/CTA повторяется последовательно или в замкнутом контуре до стабилизации аннотаций, что эмпирически улучшает точность и согласованность семантической интерпретации табличных данных [104, 78, 98].

2.1.1 Подходы к семантической интерпретации таблиц

В этом разделе рассматриваются основные подходы к Semantic Table Interpretation (STI), предложенные в научной литературе. Как отмечалось ранее, большинство исследований сосредоточено на трёх ключевых задачах – аннотации ячеек (CEA), аннотации типов столбцов (CTA) и аннотации свойств (CPA). Все существующие методы можно условно разделить на три методологических направления:

1. Эвристические методы – основаны на простых алгоритмах, таких как сопоставление строк, TF-IDF, вероятностные правила и голосование большинства.
2. Методы, основанные на инженерии признаков (Feature Engineering) – используют извлечение статистических и лексических признаков для обучения моделей машинного обучения.
3. Подходы глубокого обучения (Deep Learning) – применяют нейронные сети и представления контекста таблиц, минимизируя ручную настройку признаков.

Более подробная информация об этой классификации представлена в Таблице 2.1, где рассматриваются только задачи аннотирования, заявленные авторами, и случаи, когда проведена соответствующая оценка. Таблица 2.1 отражает эволюцию методов семантической интерпретации таблиц (STI) в рамках трёх ключевых методологических направлений: эвристических, основанных на инженерии признаков и методов глубокого обучения. Представленные подходы демонстрируют постепенный переход от правил и поисковых эвристик к более сложным итеративным и нейросетевым моделям, ориентированным на совместное решение задач CEA, CTA и CPA. Ранние эвристические методы в основном фокусировались на отдельных подзадачах, таких как определение типов столбцов или связывание сущностей, и использовали ограниченные источники знаний и пользовательские наборы данных. Итеративные подходы характеризуются более тесной интеграцией этапов аннотации таблиц и активным использованием графов знаний для повышения согласованности результатов. Методы, основанные на инженерии признаков, опираются на ручное проектирование характеристик таблиц и внешних знаний, что обеспечивает интерпретируемость, но ограничивает масштабируемость и переносимость на новые домены. В отличие от них, подходы на основе глубокого обучения обеспечивают автоматическое извлечение представлений таблиц и сущностей, что позволяет эффективно учитывать структурный, семантический и контекстный уровни данных.

Таблица 2.1 – Обзор методов по STI по трем методологическим направлениям

Подходы			CEA	CTA	CPA	Граф знаний	Источник данных	Год публикации
Класс		Алгоритм						
Эвристический	Основанный на поиске	Venetsis et al. [105]		√	√	Custom	Custom Web Tables	2011
		Wang et al. [106]		√		Probase	Custom Wikipedia Tables	2012
		Deng et al. [97]		√		FreeBase, YAGO	Custom Wikipedia Tables	2013
		Sekhavat et al. [107]			√	DBpedia	Custom Web Tables	2014
		TabEL [14]	√			YAGO	Limaye	2015
		ADOG [108]	√	√	√	DBpedia	SemTab	2019
		Tabularisi [109]	√	√	√	DBpedia	T2D, VizNet	2019
		C² [110]		√		DBpedia, Wikidata	SemTab, T2D, Custom Data	2020
		Magic [111]	√	√	√	DBpedia, Wikidata	SemTab	2021
		Alobaid et al. [112]		√		DBpedia	SemTab, T2D	2022
	Итеративный	Zwicklbauer et al. [113]		√		DBpedia	Wikipedia Tables	2013
		T2K [78]			√	DBpedia	T2D	2015
		TableMiner+ [98]			√		IMDB, MusicBrainz	2017
		LOD4ALL [100]	√	√	√	DBpedia	SemTab	2019
		CSV2KG [114]	√	√	√	DBpedia	SemTab	2019
		MTab [115][116] [117]	√	√	√	DBpedia, Wikidata	SemTab	2019
		LinkingPark [104]	√	√	√	Wikidata	SemTab	2019
		DAGOBAR SL [118][119] [99]	√	√	√	DBpedia, Wikidata	SemTab	2019
		MantisTable [120][121]	√	√	√	DBpedia, Wikidata	SemTab	2019
		JenTab [60][61][62]	√	√	√	DBpedia, Wikidata	SemTab	2020
Основанный на инженерии признаков		Limaye et al. [89]	√	√	√	YAGO	Limaye	2010
		Mulwad et al. [122][123]	√	√	√	Wikitology	Limaye	2010
		SemanticTyper [124]		√		DBpedia	Museum	2015
		DSL [125]		√		DBpedia	City, Museum, Custom Soccer	2016
		Neumaier et al. [126]		√		DBpedia	Government Data Portal	2016
		NUMER [127]		√		DBpedia	NumDB	2018

Продолжение Таблицы 2.1

На основе глубокого обучения	Моделирование графов знаний (KG)	Vasilis et al. [128]	√			Wikidata	Limaye, T2D, Wikipedia	2017
		Biswas et al. [129]				DBpedia	Custom Wikipedia infobox	2018
		DAGOBAR Embeddings [130]	√	√		DBpedia, Wikidata	SemTab	2019
		Radar Station [131]	√			Wikidata	Limaye, T2Dv2, SemTab	2022
	Моделирование таблиц	Sherlock [132]		√		DBpedia	T2D, VizNet	2019
		Sato [133]		√		DBpedia	VizNet	2019
		ColNet [134]		√		DBpedia	Limaye, T2Dv2	2019
		Guo et al. [135]		√		DBpedia	T2Dv2	2020
		Zhang et al. [80]	√		√	DBpedia	T2Dv2	2020
		TURL [101]	√	√	√	DBpedia	WikiGS, WikiTable, T2D	2020
		TCN [102]		√	√	DBpedia	Custom Web Tables, WikiTable	2021
		DUDUO [103]		√	√	-	WikiTable, VizNet	2021
		Singh et al. [136]			√	DBpedia	T2Dv2	2021
		Zhou et al. [137]		√		DBpedia	Custom Wikipedia Tables	2021

Эвристические методы объединяют широкий спектр решений, часто рассматриваемых как базовые модели STI. Они не требуют сложных этапов обучения и опираются на интуитивно понятные процедуры – строковое сходство [100,106,108], TF-IDF [108,109], голосование большинства [113] или вероятностные модели [115]. Контекст таблицы (заголовки, соседние ячейки, текстовые окружения) также учитывается, но не в полной мере [99,106]. Все подходы делятся на две подкатегории: a) методы на основе поиска (lookup-based); b) итеративные методы (iterative approaches).

Подходы на основе поиска (Lookup-Based Approaches) используют сервисы поиска для формирования набора кандидатов, а затем вычисляют оценки релевантности на основе свойств таблицы (ячейки, типы столбцов и т. д.). Venetis и др. [105] предложили модель для извлечения типов столбцов (CTA) и отношений между ключевым столбцом и другими. Для устранения ограничений существующих графов знаний авторы создали две базы: isA (для типов) и relation database (для связей), извлекая концепты из 100 миллионов текстов с помощью шаблонов вида C[such as|including]e[and|,|.]. Используя систему TextRunner [138], они построили набор отношений и применили байесовское правило и голосование большинства, показав, что гибридный

подход обеспечивает наилучшие результаты.

Итеративные методы опираются на lookup-поиск, но добавляют многошаговую дезамбигуацию для повторного ранжирования кандидатов и уточнения аннотаций, что повышает точность. Ранние подходы (Zwicklbauer и др. [113]) использовали «голосование большинства» для CTA, но резко теряли качество на малых таблицах. TableMiner/TableMiner+ [98,139] определяют столбец-субъект, итеративно аннотируют строки/столбцы с учётом HTML-контекста и partial matching; показана надёжность уже при ~8 строках. LOD4ALL [100] строит RDF-хранилище и score-БД (BM25 [140]); CTA выбирается по специфичности типа, затем CEA/CPA; близкий конвейер у CSV2KG [114] с регулировкой детализации CTA по энтропии. MTab [115] агрегирует сигналы из нескольких lookup-сервисов, заголовков, предсказаний типов и сходства значений; использует EmbNum [141] и дважды повторяет цикл CEA/CTA/CPA; версия MTab4Wikidata [116,117] улучшает опечатки и «двойной поиск», став победителем SemTab 2019–2020. LinkingPark [104] применяет каскад «coarse-to-fine» с опорой на CPA для уточнения CEA, учитывая «шум» типовой иерархии Wikidata. DAGOBAH SL [99-118] комбинирует Левенштейн и контекст узлов, использует CPA для улучшения CEA и голосование/расстояние до корня для CTA; новые версии добавили multi-hop и языковые модели и занимали 1-е места в SemTab 2021–2022. MantisTable [121,142] типизирует столбцы (NE/Literal/Subject), извлекает кандидатов по SPARQL, выполняет CEA построчно, CPA – голосованием, CTA – по частотам и иерархическим дистанциям; версия SE [120] усиливает функцию оценивания и финальную дезамбигуацию. JenTab [60-62] модульна (CEA/CTA/CPA независимы), переиспользует компоненты и показывает высокую точность на синтетике; эффект multi-hop улучшает точность до двух шагов, далее – деградация.

Подходы на основе инженерии признаков извлекают статистические и лексические признаки из строк/столбцов и обучают ML-модели (SVM, Random Forest, KNN, логистическая регрессия); качество признаков критично, а основной фокус - CTA [122,125,126]. Классические решения: TF-IDF/косинус и типовая совместимость (Limaye et al. [89]); голосование по типам + PageRank для CTA и SVM для CEA (Mulwad et al. [122]); сопоставление столбцов с ярлыками DBpedia с раздельной обработкой текстовых/числовых столбцов (SemanticTyper [124]); логистическая регрессия и случайный лес (DSL [125]); KNN с контекстом числовых столбцов и иерархией DBpedia (Neumaier et al. [126]).

Подходы на основе глубокого обучения снижает долю ручных признаков и превосходит классические ML за счёт Transformer/GNN-моделей и контекстных эмбеддингов, обучаемых совместно на таблицах и графах знаний. Два направления: (i) KG-modelling - встраивания сущностей/связей (TransE/TransH), опираясь на гипотезу семантической близости сущностей в одном столбце [128, 130, 143, 144]; (ii) Table-modelling - представление таблиц как текстово-структурированных данных с контекстными эмбеддингами для CEA/CTA/CPA.

Vasilis и соавт. [128] утверждают, что в пределах столбца «правильные»

CEA-кандидаты считаются семантически близкими. Строится взвешенный граф (веса - косинусное сходство), лучший кандидат - узел с максимальной суммой весов. Гибрид с онтологическим сопоставлением дополнительно повышает точность.

Biswas и соавт. [129] предложили метод, где тематическая аннотация инфобоксов Wikipedia по метаданным страницы (заголовки, abstract), а не по самим таблицам. Формируются Word2Vec и RDF2Vec [145] эмбеддинги, их конкатенация → классификаторы Random Forest и CNN (150 000 таблиц, 30 типов); CNN точнее RF.

DAGOBAH Embeddings [130] доказали гипотезу «сущности одного столбца близки в эмбеддинговом пространстве». Кластеризация K-Means по TransE-векторам, взвешённое голосование выбирает лучшие кластеры; одновременно извлекаются CEA и CTA. Улучшает CTA, но CEA страдает из-за ошибочных кандидатов в кластерах.

Radar Station [131] развивает DAGOBAH, добавляя семантическую дезамбигуацию после CEA; учитывает контекст столбца и графовые эмбеддинги. Даёт стабильный прирост ≈3% и показывает преимущество тонко настроенных трансляционных эмбеддингов.

Моделирование таблиц (Table Modelling) в отличие от предыдущего направления, этот подход напрямую работает с текстовым содержимым таблиц и учитывает как внутритабличные (intra-table), так и межтабличные (inter-table) взаимосвязи. Контекстуализированные представления основных элементов таблицы – ячеек, столбцов и строк – формируются при помощи глубоких нейронных сетей [134, 132] или языковых моделей, таких как BERT [101, 103, 137].

Sherlock [132] решает задачу CTA, используя 1588 признаков, извлечённых из одного столбца реляционной таблицы.

Все признаки, кроме статистических, сжимаются в фиксированные векторы через вспомогательные нейросети. Затем эти векторы и статистические признаки подаются в двухслойную полносвязную сеть, которая обучается классифицировать тип столбца среди 75 категорий набора T2Dv2. Модель показала хорошие результаты для категорий Dates и Industry, однако менее эффективно обрабатывала числовые значения и случаи, когда значения принадлежат нескольким классам одновременно.

Для решения этой проблемы Sato [133] расширил Sherlock, добавив контекст всей таблицы. Он ввёл дополнительные тематические эмбеддинги на основе LDA и слой CRF, моделирующий зависимости между столбцами.

Модель ColNet [134] ограничивается анализом одного столбца. Все слова из ячеек объединяются в последовательность, которая преобразуется в вектор Word2Vec и подаётся в CNN. Каждый столбец классифицируется независимо, без учёта межстолбцовых зависимостей. Для формирования обучающих данных используется KG, из которого извлекаются классы-кандидаты. После обучения результаты уточняются с помощью голосования большинства по внутренним ячейкам.

Guo и соавт. улучшили ColNet, заменив CNN на BiGRU-Attention и добавив Linear-CRF, что позволило учитывать зависимости между несколькими столбцами при типизации сущностей [135]. Zhang и соавт. предложили систему, совместно решающую CEA и CPA и одновременно ищущую новые сущности: она объединяет синтаксическое сходство (редакционное, Jaccard) с семантическим DRMM, а для CPA использует близость заголовков столбцов к свойствам в графе знаний [80, 146]. Модель TURL первой перенесла предобученный BERT в STI, ввела матрицу видимости для табличного контекста и, помимо CEA/CTA/CPA, решает Table Augmentation, усиливая обучение задачей Masked Entity Recovery [101, 147]. В работе Singh и соавт. отношения (CPA) извлекаются из двухстолбцовых таблиц путём линейной сериализации в BERT; обучение ведётся на синтетических таблицах из KG, где метаданные (заголовки/подписи) критично повышают качество [136]. Архитектура TCN объединяет внутри- и межтабличный контекст через механизм внимания и поддерживает transfer learning при дефиците разметки [102]. DODUO применяет многозадачное обучение: одна «голова» классифицирует тип столбца, другая предсказывает связи между столбцами; представление столбцов задаётся через токен [CLS] [103]. Наконец, Zhou и соавт. используют Star-Transformer для CTA, сочетая Word2Vec и статистические признаки по образцу Sherlock, что снижает вычислительную сложность без заметной потери точности [132, 137].

2.1.2 Методы оценки производительности систем STI

Традиционно производительность систем STI измеряется с использованием стандартных метрик информационного поиска - точности (accuracy), полноты (recall) и F1-score [130, 101, 96, 98, 125]. Среди всех задач выделяется CTA, поскольку тип и его надкласс могут одновременно считаться правильными. Например, сущность Paris может быть корректно классифицирована и как «столица», и как «город» – эти варианты не противоречат друг другу. Если считать правильным только один вариант («столица»), это усложнит объективную оценку систем, предсказывающих родовые классы.

Для решения этой проблемы в конкурсе SemTab 2019 были введены метрики AH и AP, определяемые как показаны в формулах (1-2):

$$AH = \frac{(\#PerfectA + 0.5{*}\#OkayA - \#WrongA)}{\#TargetColumns} \quad (1)$$

$$AP = \frac{\#PerfectA}{\#AnnotatedColumns} \quad (2)$$

Здесь PerfectA обозначает точное совпадение типа с эталонной разметкой, OkayA – совпадение с одним из надтипов, а WrongA – ошибочные аннотации. Такой способ оценки также используется в ColNet [134]. Конкурс SemTab Challenge проводит автоматическую оценку моделей через инструменты AIcrowd Evaluator и STILTool [148], обеспечивая стандартизированное сравнение систем по единым метрикам. В выпусках SemTab 2020 и SemTab 2021 процесс оценки задач CTA был усовершенствован введением показателя

корректности – cscore, который отражает, насколько точно предсказанный тип расположен в иерархическом дереве по отношению к эталонному (формула 3). Расстояние между предсказанным типом и истинным обозначается как d(·):

$$cscore(t)\begin{cases} 0{,}8^{d(t)} \ , if\ t\ is\ an\ ancestor\ of\ the\ GT \\ 0{,}7^{d(t)} \ , if\ t\ is\ adescendant\ og\ the\ GT \\ 0,\ otherwise; \end{cases} \quad (3)$$

Используя этот показатель, были определены приближённые метрики Precision (AP), Recall (AR) и F1-score (AF1), предназначенные для CTA (формулы 4–5):

$$AP = \frac{\Sigma cscore(t)}{\#Annotations} \qquad AR = \frac{\Sigma cscore(t)}{\#Targets} \quad (4)$$

$$AF1 = \frac{2*AP*AR}{AP+AR} \quad (5)$$

В таблице 2.2 приведены результаты трёх лучших систем по задачам CEA, CTA и CPA с наивысшими значениями F1, AP или AF1, полученными на наиболее распространённых наборах данных.

Таблица 2.2 – Результаты трёх лучших систем по задачам CEA, CTA и CPA с наивысшими значениями F1, AP или AF1, полученными на наиболее распространённых наборах данных

Датасет	CEA / Привязка строк к сущностям	CTA / Аннотация типов	CPA / Аннотация свойств
Limaye	TabEL 0.894	T2K++ 0.88	TableMiner+ 0.76
T2D	TabEAno 0.91	ColNet[e] 0.976	MantisTable 0.51
SemTab 2019	MTab 0.911	MTab 1.414	Tabularisi 0.790
SemTab 2020	MTab 0.987	JenTab 0.962	JenTab 0.963
SemTab 2021 (DBpedia)	DAGOBAH 0.945	JenTab 0.46	- -

Анализ показывает существенные различия в процедурах оценки между исследованиями. Это проявляется в следующем:

i) исходная версия набора Limaye содержала ошибки, исправленные командой TabEL, однако большинство систем STI до сих пор используют старую версию;

ii) команды TabEL и C2 публикуют только показатели точности, не предоставляя данных о полноте, что затрудняет объективное сравнение;

iii) ColNet улучшил набор T2D, но другие модели эти доработки не используют;

iv) в работе [110] представлены агрегированные результаты без

детализации по каждому набору;

v) показатели одной и той же системы на одинаковом датасете могут существенно различаться.

Например, разница в точности ColNet на наборе T2D между статьями [134] и [110] достигает ~60%, вероятно из-за отличий в настройках.

Такое несоответствие усложняет объективное сравнение систем и подчёркивает важность стандартизированных испытаний, подобных SemTab, для согласованной оценки решений в сообществе STI.

Сопоставление трёх исследовательских парадигм показало, что эвристические системы стабильно входят в тройку лидеров для всех задач и наборов данных. Ни одна модель, основанная исключительно на инженерии признаков или глубоком обучении, не вошла в топ-3 по задачам CEA или Row-to-instance. Вероятно, это связано с тем, что в отличие от CTA и CPA (где можно извлекать статистические признаки из целого столбца), аннотация отдельных ячеек или строк имеет ограниченное число признаков.

Кроме того, качество моделей обучения напрямую зависит от объёма и разнообразия кандидатов для обучения. Так как CEA требует классификацию среди миллионов сущностей, обучение становится вычислительно сложным. Поэтому в задачах CEA и row-to-instance чаще доминируют гибридные решения, совмещающие эвристику и машинное обучение.

Если рассматривать происхождение наборов данных, они делятся на синтетические (например, SemTab) и реальные веб-таблицы (Limaye, T2D). На синтетических наборах лучшие результаты показывают итеративные эвристические системы – MTab, DAGOBAH, LinkingPark, JenTab. Эти системы используют полный охват KG и эффективно оптимизируют аннотацию за счёт межзадачных связей (CEA↔CPA↔CTA). Поскольку данные таких наборов искусственно сформированы из графов знаний, строковое сопоставление работает особенно точно и не страдает от неполноты KG.

Напротив, модели, основанные на статистическом обучении (например, ColNet, Guo et al. [135]), демонстрируют лучшие результаты на компактных реальных наборах (Limaye, T2D). Это – объясняется меньшим количеством кандидатов, что упрощает обучение и повышает точность. Эвристические подходы страдают от предположения «замкнутого мира», где KG считается полным, тогда как машинное обучение и DL-модели более устойчивы к неполному покрытию графа.

За последние годы область Semantic Table Interpretation (STI) претерпела значительный рост благодаря разработке новых методов и появлению масштабных наборов данных, в частности в рамках инициативы SemTab Challenge.

В данном обзоре представлен комплексный и актуальный анализ состояния исследований в STI. Работа систематизирует существующие определения, описывает входные данные процесса – типы таблиц, их контекст и метаданные – и классифицирует подходы в зависимости от используемой парадигмы: эвристической, фичерной (feature engineering) и нейросетевой (deep

learning).

Также рассмотрены ключевые сильные и слабые стороны каждого подхода, проведено сравнение производительности различных систем на популярных наборах данных и выделены лучшие решения для каждой задачи.

Наконец, определены основные вызовы и перспективы развития, включая расширение типов таблиц, преодоление неполноты графов знаний и интеграцию более богатого контекста. Современные тенденции также включают создание интерактивных веб-интерфейсов (например, [87]), что сделает технологии STI более доступными и понятными для конечных пользователей.

2.2 Сбор данных из гетерогенных источников по водному домену Казахстана

2.2.1 Область исследования

В качестве предметной области исследования в настоящей работе рассматриваются водные ресурсы Республики Казахстан. Поскольку к водным объектам относятся различные типы источников – реки, озёра, каналы, водохранилища – и они распределены по всей территории страны, на первоначальном этапе планировалось сосредоточить внимание на водных ресурсах Алматинской области. Однако в процессе изучения официальных гидрологических материалов было установлено, что классификация водных объектов в Казахстане базируется не на административно-территориальном делении, а на бассейновом принципе.

Согласно данным РГП «Казгидромет», представленным в разделе «Гидрология» официального сайта и ежегодном издании *«Поверхностные воды* [149]. Ежегодные данные о режиме и ресурсах поверхностных вод суши», водные объекты Казахстана структурированы по восьми гидрографическим бассейнам:

1. бассейн реки Ертис;
2. бассейн реки Есиль;
3. бассейны рек Тобол и Торгай;
4. бассейн реки Урал;
5. бассейн реки Сырдария;
6. бассейны рек Шу и Талас;
7. бассейны рек озера Балхаш и озера Алаколь;
8. бассейны рек Нура и Сарысу.

Границы территорий водохозяйственных бассейнов Республики Казахстан, указаны на Рисунке 2.8.



1 – границы водохозяйственных бассейнов; 2 – границы административных областей

Рисунок 2.8 - Схема деления водохозяйственных бассейнов Республики Казахстан

В рамках данного исследования в качестве объекта анализа выбран бассейн рек озера Балхаш и озера Алаколь. Сбор и обработка исходных данных выполнялись на основе материалов соответствующего тома ежегодника «Поверхностные воды».

В ходе дополнительного анализа установлено, что бассейн озера Балхаш в научной и справочной литературе, а также в энциклопедических источниках (в частности, в Википедии) часто упоминается под названием Или-Балхашский бассейн (ИББ) [150]. Данное обозначение отражает ключевую роль реки Или как главного водотока бассейна.

В настоящей работе для обеспечения терминологической согласованности используется название «Или-Балхашский бассейн» (сокращённо – ИББ). При этом в анализ также включаются данные РГП «Казгидромет» по бассейну озера Балхаш и озера Алаколь, поскольку состав гидросети в обоих случаях охватывает идентичные или перекрывающиеся водные объекты.

Или-Балхашский бассейн (ИББ) является частью Балкаш-Алакольского бассейна и представляет собой одну из крупнейших гидрографических систем Центральной Азии, охватывающую трансграничный водосбор, расположенный на территории Республики Казахстан и Китайской Народной Республики. Общая площадь Балкаш-Алакольского бассейна превышает 500 тыс. км², при этом площадь водосборного бассейна непосредственно озера Балхаш составляет около 413 тыс. км², из которых 15 % находятся в Синьцзян-Уйгурском автономном районе КНР, а оставшаяся территория – в Казахстане (≈85 %). Среднегодовой суммарный поверхностный сток бассейна оценивается в 28,0 км³, из которых 11,8 км³ поступает с территории Китая. Более 90 % всех рек бассейна относятся к водосбору озера Балхаш. Главным водотоком является река Или, формирующая около 75 % суммарного притока в озеро; её питание преимущественно ледниковое, что обуславливает выраженную сезонную

динамику уровня воды с пиком в июне–июле. Восточная часть озера питается водами рек Каратал, Аксу, Лепсы, а также подземными водами [151].

Бассейн охватывает обширные территории Алматинской, Жамбылской, Карагандинской и Восточно-Казахстанской областей, включая ряд городских агломераций (г. Балхаш, г. Приозерск, часть г. Алматы) и значительные сельскохозяйственные угодья. На востоке и юго-востоке границы ИББ определяются горными системами Джунгарского Алатау, Терскей и Кунгей Алатау, Заилийского Алатау и Чу-Илийских гор. Географические координаты бассейна варьируются в пределах 42°–47° с. ш. и 74°–85° в. д., что обусловливает его выраженное внутриконтинентальное положение и формирует специфические климатические условия.

Согласно данным РГП «Казгидромет», в пределах Или-Балхашского бассейна функционируют 69 гидрологических постов. На рисунке 2.9 представлена схема расположения гидрологических постов ИББ.

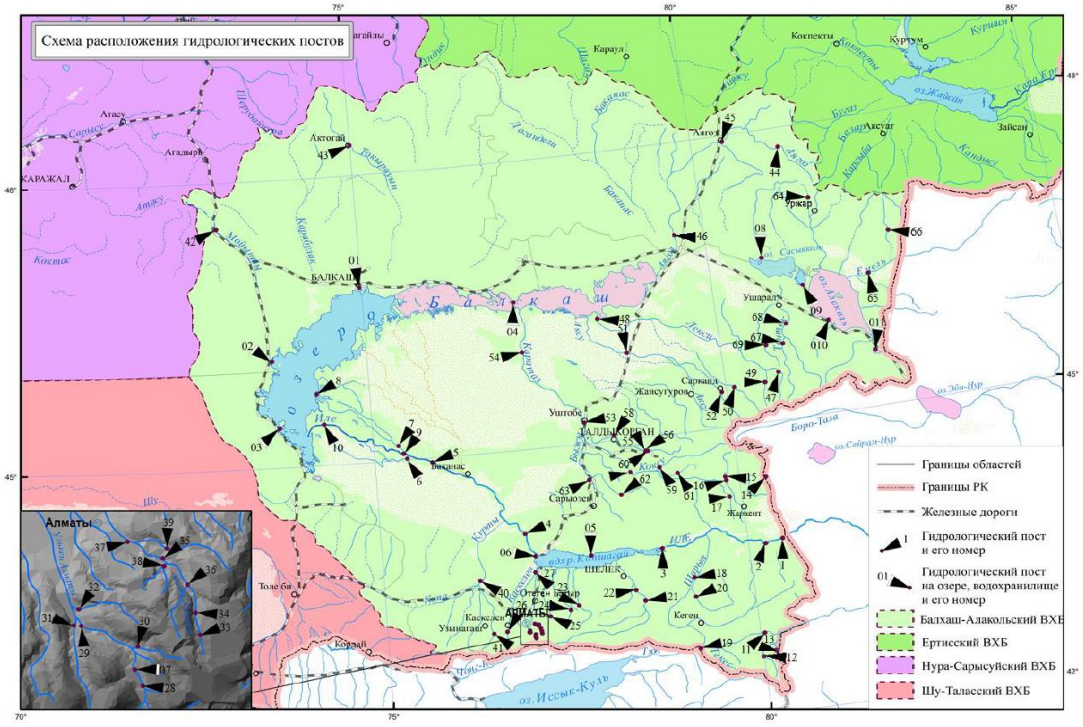


Рисунок 2.9 – Схема расположения гидрологических постов Или-Балхашского бассейна [152]

В терминологии гидрологического мониторинга гидрологический пост представляет собой пункт наблюдений, расположенный на водном объекте и оснащённый специализированными приборами и устройствами для проведения систематических измерений гидрологических параметров. Каждому посту присвоен уникальный код, который используется в запросах к архивным материалам РГП «Казгидромет».

Существенное влияние на гидрологический режим оказывает зарегулирование стока. В 1970 г. на реке Или была введена в эксплуатацию Капшагайская ГЭС мощностью 364 МВт, создавшая водохранилище,

используемое как для выработки электроэнергии, так и для целей орошения. В китайской части бассейна в 2005–2006 гг. введены в строй водохранилища на реках Текес и Каш, суммарный объём которых превышает 4,5 км³.

Гидрологическое районирование бассейна выделяет несколько крупных зон:

- бассейн реки Или (включая бассейны рек Илийского Алатау и хребта Кетмень);
- бассейны рек Жетысуского Алатау;
- Северное Прибалкашье (включая бассейны рек Мойынты, Токырауын и южных склонов хребта Тарбагатай).

В ежегодниках РГП «Казгидромет» приводятся сведения о сезонной динамике стока, уровнях и расходах воды, а также об особенностях гидрологического режима по результатам наблюдений за гидрологический год (с 1 октября предыдущего года по 30 сентября отчётного).

2.2.2 Гетерогенные источники по водному домену Казахстана

Мониторинг водных ресурсов Республики Казахстан опирается на массив разнотипных данных, поступающих из гетерогенных источников и представленных в неоднородных форматах. Для корректной оценки состояния водных экосистем необходимы:

(ii) химико-аналитические показатели качества воды (pH, растворённый кислород, БПК, нутриенты, металлы и др.);

(iii) инструментальные ряды наблюдений гидрологических постов (уровни, расходы, температуры, ледовые явления);

(iv) метаданные об объектах и пунктах наблюдений (местоположение, система высот, период действия, ведомственная принадлежность);

(v) нормативно-правовые критерии интерпретации измерений (классы качества, предельно допустимые концентрации).

Учитывая юрисдикционную специфику нормирования, в первую очередь для водных объектов РК были собраны данные по «Водным правилам». Нормативной основой послужил официальный документ: Приказ Министра водных ресурсов и ирригации Республики Казахстан от 4 июня 2025 г. № 111-НҚ «Об утверждении единой системы классификации качества воды в поверхностных водных объектах и (или) их частях» [153]. На рисунке 2.10 приведена главная страница сайта информационно-правовой системы нормативных правовых актов Республики Казахстан «Әділет».

На основании данного акта выделены и приведены к табличной форме следующие наборы данных, предназначенные для последующего пополнения Модуля «Водные правила»:

1. Единая система классификации воды в поверхностных водных объектах и (или) их частях – перечень нормируемых показателей качества, единицы измерения и интервальные критерии для отнесения воды к классам 1–6.

2. Характеристика классов качества воды – описательные

характеристики каждого класса (экологическое состояние, пригодность использования, признаки деградации), используемые для текстовой интерпретации результатов.

3. Дифференциация классов по категориям (видам) водопользования – правила допустимости классов качества для различных целей (хозяйственно-питьевое, коммунально-бытовое, рыбохозяйственное и др.).

4. ПДК загрязняющих веществ – предельно допустимые концентрации химических веществ/показателей в разрезе категорий водопользования с указанием величин и единиц.

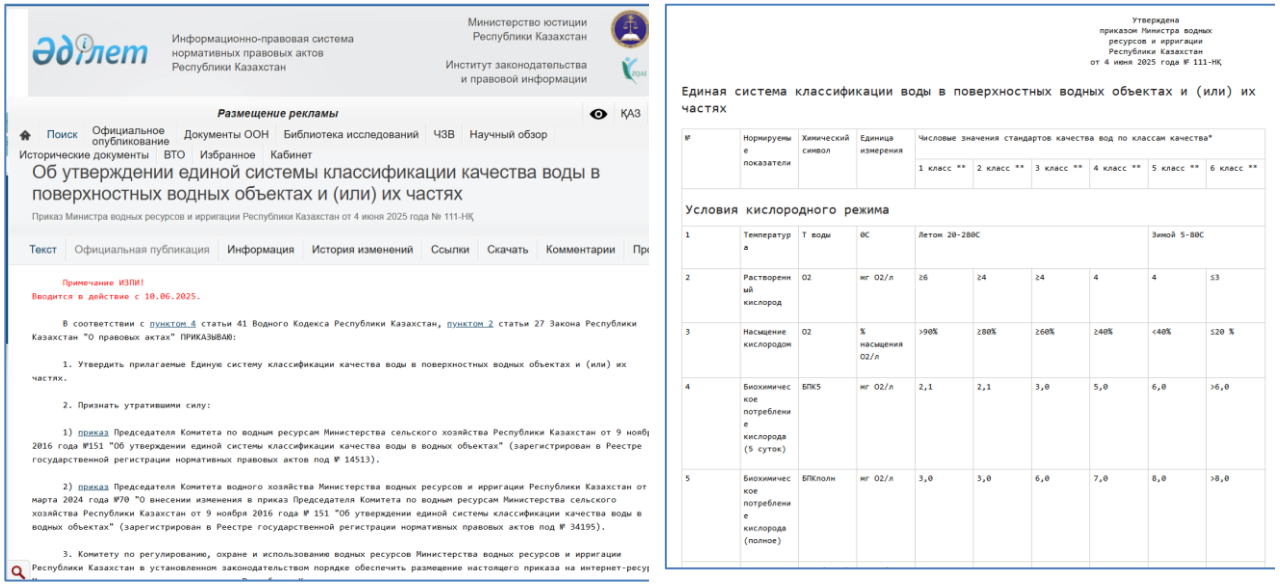


Рисунок 2.10 – Веб-страница информационно-правовой системы нормативных правовых актов о водном кодексе Республики Казахстан «Әділет»

В качестве второго источника использованы материалы раздела «Гидрология» официального сайта РГП «Казгидромет» – серия Государственного водного кадастра «Поверхностные воды: Ежегодные данные о режиме и ресурсах поверхностных вод суши (ЕДС)» за 2002–2023 гг. по бассейнам оз. Балкаш и оз. Алаколь [154]. Публикации представлены в формате PDF и содержат сведения о сети гидрологических постов и результатах наблюдений (уровень, расход, режимные характеристики), а также служебные параметры (коды объектов, отметки нулей постов, системы высот, периоды действия и пр.) (см.Рис.2.11).

На основании указанных ежегодников подготовлены и приведены к машиночитаемому виду следующие табличные наборы данных для последующего использования в модуле Hydroposts:

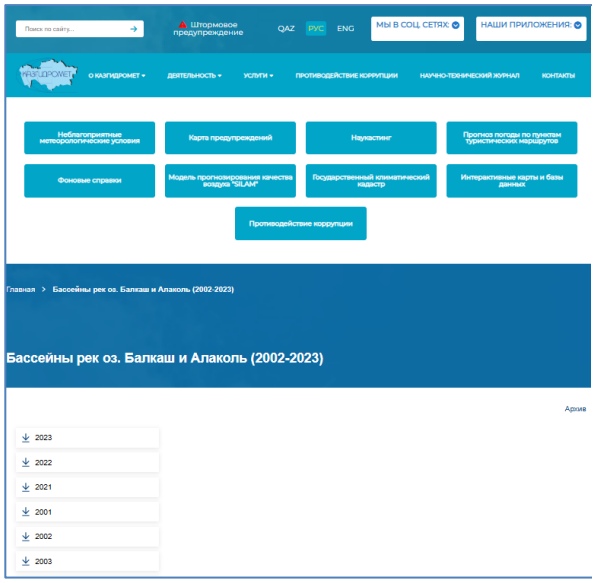
1. Список постов на реках и каналах (ИББ) – таблица состоит из следующих полей: наименование водного объекта (участок/привязка); код водного объекта; код поста; расстояние от устья, км; площадь водосбора, км²; отметка нуля, м; система высот (Балтийская система – «БС» или условная – «усл.»); дата открытия; при наличии – доп. дата (перенос/изменение); статус; принадлежность. Таблица консолидирует паспортные сведения по сети

речных/канальных гидропостов в границах Иле-Балхашского бассейна.

2. Список постов на озёрах и водохранилищах (ИББ) – таблица состоит из следующих полей: местоположение поста; код водного объекта; код поста; площадь водосбора, км²; площадь зеркала водоёма, км²; отметка нуля, м; система высот; дата открытия; при наличии – дата закрытия и/или доп. дата (перенос/изменение); статус; принадлежность поста. Таблица охватывает стационарные пункты наблюдений на озёрах и водохранилищах бассейна.

3. Месячные ряды уровня воды (ИББ) – таблица состоит из следующих полей: Код_поста; Дата (в формате YYYY-MM); Уровень_воды_Значение; Бассейн (значение: BALHASH-ALAKOL). Набор содержит агрегированные по месяцам значения уровня воды для соответствующих гидропостов (в единицах, принятых в первоисточнике ежегодника), с сохранением официального кода поста и принадлежности к бассейну.

4. Месячные ряды расхода воды (ИББ) – таблица состоит из следующих полей: Код_поста; Дата (формат YYYY-MM); Расход_воды_Значение. Представлены агрегированные по месяцам значения расхода воды (в единицах, указанных в ежегодниках), синхронизированные по тем же кодам постов, что и в списках сети.



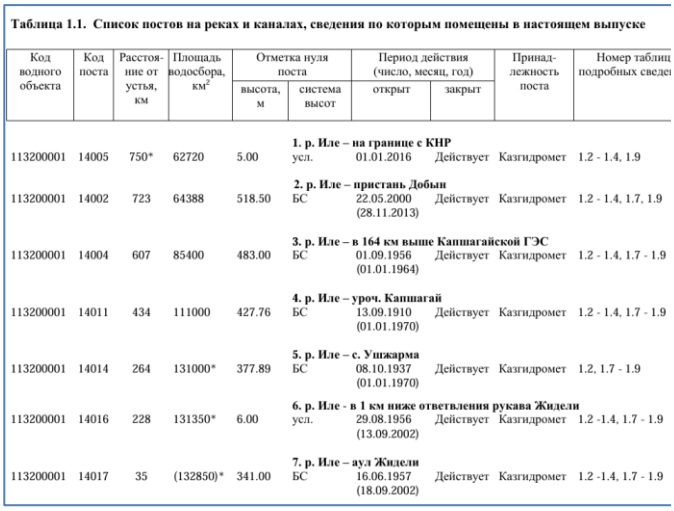


Таблица 1.1. Список постов на реках и каналах, сведения по которым помещены в настоящем выпуске

Код водного объекта	Код поста	Расстояние от устья, км	Площадь водосбора, км²	Отметка нуля поста		Период действия (число, месяц, год)		Принадлежность поста	Номер таблиц подробных сведе
				высота, м	система высот	открыт	закрыт		
					1. р. Иле – на границе с КНР				
113200001	14005	750*	62720	5.00	усл.	01.01.2016	Действует	Казгидромет	1.2 - 1.4, 1.9
					2. р. Иле – пристань Добын				
113200001	14002	723	64388	518.50	БС	22.05.2000 (28.11.2013)	Действует	Казгидромет	1.2 - 1.4, 1.7, 1.9
					3. р. Иле – в 164 км выше Капшагайской ГЭС				
113200001	14004	607	85400	483.00	БС	01.09.1956 (01.01.1964)	Действует	Казгидромет	1.2 - 1.4, 1.7 - 1.9
					4. р. Иле – уроч. Капшагай				
113200001	14011	434	111000	427.76	БС	13.09.1910 (01.01.1970)	Действует	Казгидромет	1.2 - 1.4, 1.7 - 1.9
					5. р. Иле – с. Ушжарма				
113200001	14014	264	131000*	377.89	БС	08.10.1937 (01.01.1970)	Действует	Казгидромет	1.2, 1.7 - 1.9
					6. р. Иле - в 1 км ниже ответвления рукава Жидели				
113200001	14016	228	131350*	6.00	усл.	29.08.1956 (13.09.2002)	Действует	Казгидромет	1.2 -1.4, 1.7 - 1.9
					7. р. Иле – аул Жидели				
113200001	14017	35	(132850)*	341.00	БС	16.06.1957 (18.09.2002)	Действует	Казгидромет	1.2 -1.4, 1.7 - 1.9

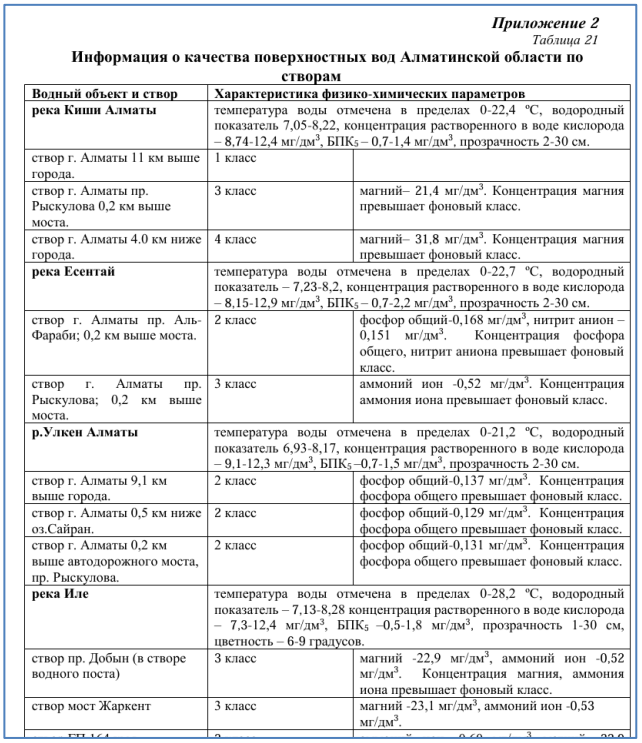
Рисунок 2.11 – Официальный сайт РГП «Казгидромет», серия Государственного водного кадастра «Поверхностные воды»

Во всех таблицах сохранены официальные идентификаторы (код водного объекта, код поста), принятая в ежегодниках система высот (БС/усл.), а также календарная привязка наблюдений. Даты во временных рядах нормализованы к формату YYYY-MM для удобства последующего анализа и сопоставления.

Также для модуля Гидропостов надо было собрать данные о наблюдениях по химическому составу водных объектов. В рамках раздела «Экология» официального сайта РГП «Казгидромет» публикуются ежемесячные информационные бюллетени о состоянии окружающей среды по регионам Республики Казахстан. Бюллетени выходят до 17-го числа месяца, следующего

за отчётным, с выбором года и поквартальной/помесячной навигацией по субъектам (в т. ч. г. Алматы, Алматинская область и др.) [155]. Эти издания содержат структурированные разделы, среди которых (в разных выпусках и регионах) встречаются: результаты мониторинга качества поверхностных вод, состояние атмосферного воздуха, химический состав атмосферных осадков, состояние донных отложений, загрязнение почв, радиационная обстановка и приложения с методическими/справочными сведениями (Рис.2.12).





Приложение 2

Таблица 21

Информация о качества поверхностных вод Алматинской области по створам

Водный объект и створ	**Характеристика физико-химических параметров**	
река Киши Алматы	температура воды отмечена в пределах 0-22,4 ºС, водородный показатель 7,05-8,22, концентрация растворенного в воде кислорода – 8,74-12,4 мг/дм^3, БПК$_5$ – 0,7-1,4 мг/дм^3, прозрачность 2-30 см.	
створ г. Алматы 11 км выше города.	1 класс	
створ г. Алматы пр. Рыскулова 0,2 км выше моста.	3 класс	магний– 21,4 мг/дм^3. Концентрация магния превышает фоновый класс.
створ г. Алматы 4.0 км ниже города.	4 класс	магний– 31,8 мг/дм^3. Концентрация магния превышает фоновый класс.
река Есентай	температура воды отмечена в пределах 0-22,7 ºС, водородный показатель – 7,23-8,2, концентрация растворенного в воде кислорода – 8,15-12,9 мг/дм^3, БПК$_5$ – 0,7-2,2 мг/дм^3, прозрачность 2-30 см.	
створ г. Алматы пр. Аль-Фараби; 0,2 км выше моста.	2 класс	фосфор общий-0,168 мг/дм^3, нитрит анион – 0,151 мг/дм^3. Концентрация фосфора общего, нитрит аниона превышает фоновый класс.
створ г. Алматы пр. Рыскулова; 0,2 км выше моста.	3 класс	аммоний ион -0,52 мг/дм^3. Концентрация аммония иона превышает фоновый класс.
р.Улкен Алматы	температура воды отмечена в пределах 0-21,2 ºС, водородный показатель 6,93-8,17, концентрация растворенного в воде кислорода – 9,1-12,3 мг/дм^3, БПК$_5$ –0,7-1,5 мг/дм^3, прозрачность 2-30 см.	
створ г. Алматы 9,1 км выше города.	2 класс	фосфор общий-0,137 мг/дм^3. Концентрация фосфора общего превышает фоновый класс.
створ г. Алматы 0,5 км ниже оз.Сайран.	2 класс	фосфор общий-0,129 мг/дм^3. Концентрация фосфора общего превышает фоновый класс.
створ г. Алматы 0,2 км выше автодорожного моста, пр. Рыскулова.	2 класс	фосфор общий-0,131 мг/дм^3. Концентрация фосфора общего превышает фоновый класс.
река Иле	температура воды отмечена в пределах 0-28,2 ºС, водородный показатель – 7,13-8,28 концентрация растворенного в воде кислорода – 7,3-12,4 мг/дм^3, БПК$_5$ –0,5-1,8 мг/дм^3, прозрачность 1-30 см, цветность – 6-9 градусов.	
створ пр. Добын (в створе водного поста)	3 класс	магний -22,9 мг/дм^3, аммоний ион -0,52 мг/дм^3. Концентрация магния, аммония иона превышает фоновый класс.
створ мост Жаркент	3 класс	магний -23,1 мг/дм^3, аммоний ион -0,53 мг/дм^3.

Рисунок 2.12 – Информационный бюллетень о состоянии окружающей среды по регионам Республики Казахстан на сайте РГП «Казгидромет»

Наличие и нумерация разделов подтверждаются типовыми PDF-выпусками на сайте Казгидромета (и зеркалах госфондов), где явно присутствуют подразделы «Состояние/результаты мониторинга качества поверхностных вод».

В настоящем исследовании из этих бюллетеней целенаправленно извлекались только таблицы, относящиеся к состоянию качества поверхностных вод, причём по г. Алматы и Алматинской области (основной регион расположения Иле-Балхашского бассейна).

Эти файлы представляют месячные/квартальные выборки табличных данных по мониторингу поверхностных вод для выбранных территорий за соответствующие периоды. Структура полей следует исходным таблицам бюллетеней и, как правило, включает идентификаторы пункта наблюдений и водного объекта, перечень определяемых показателей качества и их измеренные значения с единицами (в форме, принятой в первоисточнике), а также служебные примечания/ссылки (при наличии). Наборы предназначены для использования в модуле, посвящённом сети гидропостов и результатам наблюдений.

Также одним из ключевых источников для получения гидрологических параметров по водным объектам Иле-Балхашского бассейна стала “Гидрологическая база данных Национальной гидрометеорологической службы РГП «Казгидромет»”, доступная в открытом доступе через веб-приложение по адресу: http://ecodata.kz:3838/app_hydro/.

Данная база данных разработана на основании Правил предоставления информации Национальной гидрометеорологической службой, утвержденных приказом Министра экологии, геологии и природных ресурсов Республики Казахстан №267 от 23 июля 2021 года (рег. №23716 от 27 июля 2021 года). В ней представлены режимные гидрологические наблюдения, полученные на постах государственной наблюдательной сети с использованием автоматизированной системы первичной обработки данных «Реки-Режим». На Рис. 2.13 представлен фрагмент пользовательского интерфейса официального веб-приложения Гидрологической базы данных, разработанного РГП «Казгидромет».

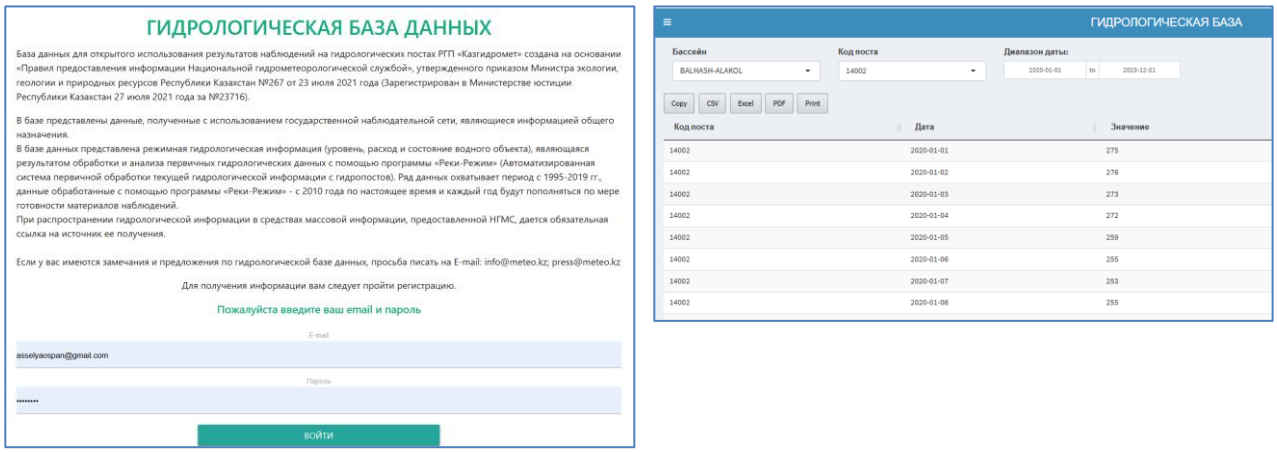


Рисунок 2.13 - Веб-интерфейс базы данных гидрологических наблюдений «Реки-Режим» (ecodata.kz)

Для целей диссертационного исследования были использованы данные об уровнях и расходах воды для водотоков Балхаш-Алакольского бассейна, охватывающие период с 2020 по 2022 годы, а также оперативные измерения, обрабатываемые в режиме ежегодного пополнения. Наблюдения охватывают более 60 гидрологических постов.

Информация включает следующие параметры:

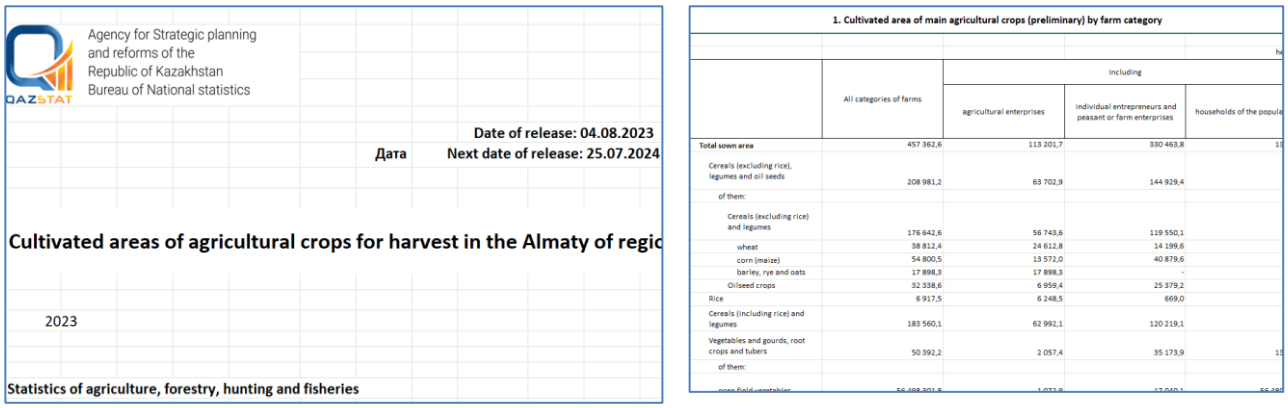
- Уровень воды – высота водной поверхности относительно реперной отметки (измеряется в см), рассчитываются среднесуточные и среднемесячные значения.
- Расход воды – объём воды, протекающий через живое сечение потока в единицу времени (м³/с), с вычислением среднемесячных и экстремальных значений.

Данные проходят контроль качества и агрегируются в виде таблиц, соответствующих формату официальных ежегодников: «Ежегодные данные о режиме и ресурсах поверхностных вод суши» (ЕДС), которые используются в

работе для построения семантических триплетов и последующего пополнения онтологической модели.

Применение данной базы позволило обеспечить семантически выровненное и стандартизированное представление гидрологических характеристик, интегрированных в единую онтологию мониторинга водных ресурсов Казахстана.

Для построения графа знаний по водным ресурсам дополнительно использованы таблицы со stat.gov.kz по региональным показателям: численность населения, рождаемость/смертность, коэффициенты естественного движения, число промышленных предприятий (по ОКЭД), площади сельскохозяйственных угодий (в т.ч. пашни). Эти данные интегрированы с водными объектами ИББ через региональную привязку и используются для последующего анализа связей «состояние вод» ↔ «социально-экономические факторы». На рисунке 2.7 приведена таблица площади пахотных земель Алматинской области. В целом были загружены таблицы по этому показателю для всех областей, входящих в бассейн Иле-Балхаш (Алматинская, Жетысуская, Карагандинская, ВКО).


Agency for Strategic planning and reforms of the Republic of Kazakhstan Bureau of National statistics

QAZSTAT

Date of release: 04.08.2023

Дата Next date of release: 25.07.2024

Cultivated areas of agricultural crops for harvest in the Almaty of regic

2023

Statistics of agriculture, forestry, hunting and fisheries

1. Cultivated area of main agricultural crops (preliminary) by farm category

	All categories of farms	Including		
		agricultural enterprises	individual entrepreneurs and peasant or farm enterprises	households of the popula
Total sown area	457 362,6	113 201,7	330 463,8	1
Cereals (excluding rice), legumes and oil seeds	208 981,2	63 702,9	144 929,4	
of them:				
Cereals (excluding rice) and legumes	176 642,6	56 743,6	119 550,1	
wheat	38 812,4	24 612,8	14 199,6	
corn (maize)	54 800,5	13 572,0	40 879,6	
barley, rye and oats	17 898,3	17 898,3	-	
Oilseed crops	32 338,6	6 959,4	25 379,2	
Rice	6 917,5	6 248,5	669,0	
Cereals (including rice) and legumes	183 560,1	62 992,1	120 219,1	
Vegetables and gourds, root crops and tubers	50 392,2	2 057,4	35 173,9	1
of them:				
open field vegetables	56 498 201,8	1 073,8	17 049,1	56 480

Рисунок 2.14 – Площадь пахотных земель сельскохозяйственной культуры Алматинской области

Таким образом, по водному домену Иле-Балхашского бассейна были собраны 25 таблиц. Далее эти таблицы будут использованы для пополнения графов знаний.

2.3 Инструменты извлечения таблиц из неструктурированных источников

В разделе рассматриваются практические средства извлечения табличных данных из PDF, HTML и полуструктурированных Excel/CSV: QURMA для веб-страниц, TableProcessor для “грязных” офисных таблиц и PyMuPDF (в связке с Camelot/Tabula при необходимости) для PDF. Описываются форматы входа/выхода, правила нормализации, обработка заголовков, объединённых ячеек и типизация столбцов для дальнейшей семантической интерпретации (STI).

PyMuPDF [156] – это высокопроизводительная Python-обёртка над движком MuPDF, предназначенная для чтения, разбора, поиска, извлечения текста, графики и разметки из PDF и ряда других форматов (XPS, EPUB, CBZ и

др.). Библиотека предоставляет удобный API уровня страниц и объектов (шрифты, аннотации, изображения), а также методы для точного позиционирования текста (слова/строки/спаны) и извлечения фрагментов по прямоугольным областям, что делает её особенно подходящей для задач предварительной обработки табличных данных перед семантической интерпретацией. Официальная документация содержит примеры получения текстовых “слов” с координатами (page.get_text("words")), поиска по шаблонам и подсветки найденного текста, экспорта в HTML/JSON, извлечения и повторного встраивания изображений, а также работы с аннотациями и формами.

PyMuPDF активно применяется в исследовательских и прикладных проектах по интеллектуальной обработке документов (Document AI). Например, инженерная команда Artifex (поддерживающая MuPDF) демонстрирует конвейеры распознавания и извлечения таблиц на базе MuPDF с интеграцией ML-моделей для детекции табличных структур, где движок обеспечивает быстрый доступ к графическим примитивам и текстовым спанам, необходимым для последующей сегментации макета и восстановления границ ячеек [157]. Эти материалы полезны как практические руководства по построению пайплайнов «PDF → структурированные представления → табличные данные».

В академической среде PyMuPDF используется как стандартный слой загрузки и стратификации PDF при создании мультимодальных парсеров. Так, в недавней работе по построению многоцелевого парсера документов (мультимодальная сегментация и извлечение структур) PyMuPDF применяется для итерации по страницам, извлечения текстовых блоков и визуальных фич, после чего результаты скармливаются нейросетевым компонентам, обученным на задачах разметки макета и извлечения сущностей. Авторы подчёркивают важность точных координат и воспроизводимого порядка обхода объектов, которые предоставляет API библиотеки.

PyMuPDF часто выбирают и в индустриальных решениях для построения «легковесных» ETL-контуров поверх PDF, где требуется массовое извлечение текстов/таблиц, конвертация в JSON/CSV и индексация [158]. Примером служат платформы документ-интеллекта, где PyMuPDF используется как быстрый и минималистичный экстрактор перед подачей данных в OCR/LLM компоненты и индексаторы: типичный сценарий – извлечь страницы, текстовые блоки и медиа, нормализовать их и сопоставить с доменными онтологиями.

В рамках диссертации QURMA [159] используется как лёгкий модуль веб-скрейпинга и табличной нормализации для получения регуляторных данных из нормативных источников (в частности, с портала zan.kz, где опубликованы гигиенические нормативы за качеством воды). Практически QURMA автоматизирует: (i) загрузку HTML-страниц, (ii) детекцию и извлечение HTML-таблиц, (iii) очистку/нормализацию (слияние многоуровневых заголовков, разворачивание «rowspan/colspan», унификацию единиц измерения, раскодировку аббревиатур), (iv) экспорт в единый табличный формат (CSV/Excel) для последующей семантической интерпретации. Такой подход

соотносится с классической линией работ по извлечению и осмыслению веб-таблиц, где таблицы рассматриваются как богатый источник структурированных фактов для построения графов знаний и аннотирования сущностей [160].

Технически QURMA опирается на стандартный стек: HTTP-клиент для стабильной загрузки страниц, BeautifulSoup/bs4 для парсинга и очистки DOM, а также высокоуровневые рутинги наподобие pandas.read_html для первичного считывания табличных узлов и конвертации их в DataFrame с последующей нормализацией [161]. Такой гибридный дизайн позволяет сочетать эвристики макетной разметки (layout-aware) с детерминированным парсингом HTML-таблиц, как это рекомендовано в работах по восстановлению «логической» структуры таблиц, где важно различать декоративные таблицы и «подлинные», несущие реляционное содержание.

Содержательно в вашем проекте QURMA применялся для выгрузки и унификации таблиц гигиенических нормативов (перечень регулируемых показателей: растворённый кислород, БПК, ХПК, нитраты/нитриты, тяжёлые металлы, ПАВ и т.п.; пороговые значения для классов качества воды; единицы измерения). Эти таблицы затем поступают в модуль Semantic Table Interpretation (STI), где маппируются на словари/онтологии (например, показатели – к справочнику индикаторов качества, единицы – к онтологиям единиц), а пороговые значения – к правилам SWRL. Такой переход от «сырых» таблиц к семантически размеченным наборам согласуется с современными практиками построения графов знаний из веб-таблиц и их последующей публикацией в формате RDF/Linked Data [162].

Практические преимущества QURMA заключаются в: (1) воспроизводимости (декларативные правила нормализации и журналирование парсинга), (2) масштабируемости на новые страницы/разделы без изменения ядра, (3) совместимости с последующими стадиями – от OpenRefine/CSV до прямой сериализации в RDF, что соответствует проверенным маршрутам «таблица → KG» в литературе. В сумме это позволяет использовать QURMA как «входной шлюз»: от нормативной веб-публикации к чистой, унифицированной таблице, готовой к семантической интерпретации и интеграции в онтологическую модель мониторинга водных ресурсов.

Для извлечения структурных данных из ексель таблиц, в рамках данного исследования был разработан инструмент TableProcessor для интерпретации Excel-таблиц. TableProcessor [163] реализует практический конвейер STI для Excel-таблиц со статистическими и социально-экономическими показателями. Мы придерживаемся класса таблиц, где один столбец содержит имена сущностей (ключевой/subject), остальные – их атрибуты/значения. Требования к данным: наличие заголовков, отсутствие объединений, не менее одной текстовой колонки-сущности и ≥10 строк. Инструмент автоматически: (i) нормализует структуру, (ii) детектирует subject-столбец (эвристика «самый левый с наибольшей долей уникальных нечисловых значений»), (iii) формирует кандидатов сущностей (по лексическим индексам KG), (iv) вычисляет приоритеты/достоверности сопоставлений с учётом контекста (заголовки

таблицы/страницы, соседние столбцы/строки), (v) выполняет дезамбигуацию (итеративно, до сходимости), (vi) извлекает тройки (\langle s,r,o\rangle) и сохраняет результат в JSON/RDF для дальнейшей загрузки в KG. Для ускорения мы применяем выборку ячеек на этапе предклассификации столбцов (с критерием остановки), что позволяет избежать исчерпывающей обработки без потерь точности, сопоставимой с «полным» проходом. Отношения между subject-столбцом и сущностными столбцами (CPA) выводятся из кандидатных пар с учётом онтологического и табличного контекста, а литеральные столбцы маркируются как data properties. В итоге формируется компактное множество триплетов, готовых к загрузке в граф знаний и последующему семантическому выводу.

Источником Excel-таблиц послужил открытый портал Национального бюро статистики (stat.gov.kz) с показателями по регионам РК; из массива >1000 таблиц отобраны тематические наборы (труд/занятость, уровень жизни) и нормализованы под фиксированную схему для тестирования. TableProcessor обрабатывает пакет файлов по пути-каталогу и генерирует для каждого исходника парный JSON с авто-выделением заголовка, subject-столбца, атрибутов, значений и сопроводительного текста. Эти JSON-файлы далее используются в ГИС и в модуле построения KG.

3 МЕТОДОЛОГИЯ ФОРМИРОВАНИЯ ГЕОИНФОРМАЦИОННОЙ МОДЕЛИ НА ОСНОВЕ ОНТОЛОГИЙ ДЛЯ ИЗВЛЕЧЕНИЯ ЗНАНИЙ

Целью настоящего раздела является разработка и формализация целостной методологии построения онтологически-ориентированной геоинформационной модели, обеспечивающей автоматизированное извлечение, семантическую унификацию, интеграцию и логический вывод знаний из гетерогенных источников данных о водных ресурсах. Предлагаемая методология направлена на формирование воспроизводимого графа знаний, объединяющего пространственные, временные, измерительные и нормативные аспекты мониторинга в рамках единой семантической архитектуры. Тем самым обеспечивается переход от разрозненных табличных и текстовых представлений к структурированной интеллектуальной системе, способной поддерживать комплексный пространственно-временной анализ состояния водных объектов и выявление закономерностей их изменения.

В данном разделе формируется архитектурный и алгоритмический каркас системы, основанный на принципах многоуровневой обработки данных, онтологического выравнивания (ontology alignment) и семантического вывода. Разрабатывается модульная структура гидроонтологии, предусматривающая согласование с международными стандартами (WHOW/SSN/SOSA, HY_Features, OWL-Time) и их расширение национальными доменными модулями, отражающими нормативные и социально-географические особенности мониторинга водных ресурсов Казахстана. Особое внимание уделяется формализации итеративного метода семантической интерпретации табличных данных, обеспечивающего автоматическое выявление субъектных и атрибутивных структур, нормализацию единиц измерения, построение и валидацию RDF-триплетов с сохранением метаданных происхождения и временной привязки наблюдений. Дополнительно разрабатываются правила логического вывода на языке SWRL, обеспечивающие автоматическую классификацию качества воды на основе нормативных порогов, связывание наблюдений с пространственными и временными контекстами и индуктивное выявление новых семантических зависимостей.

Структура дальнейшего изложения главы выстроена в соответствии с логикой трёхуровневой архитектуры и направлена на последовательное раскрытие её компонентов – от концептуального проектирования системы до механизмов автоматизированного логического вывода и геопространственной визуализации результатов. Такая композиция обеспечивает системность представления разработанной методологии и отражает поэтапный переход от извлечения и структурирования данных к формированию интегрированной интеллектуальной геоинформационной среды.

В пункте 3.1 рассматривается многоуровневая архитектура онтологически-ориентированной геоинформационной модели для извлечения знаний, где обосновывается структурная декомпозиция системы на уровни семантической интерпретации данных, онтологической унификации и генерации результатов. В пункте 3.2 анализируется модульная архитектура гидроонтологии, включающая

описание пространств имён, ключевых классов и свойств международного слоя, а также концептуальное обоснование и структуру национальных модулей WR_KZ и SocioGeo_KZ. Особое внимание уделяется механизмам онтологического выравнивания и повторного использования терминологии, обеспечивающим согласованность с международными онтологическими моделями и достижение семантической интероперабельности.

В пункте 3.3 представлена формализация итеративного STI-конвейера, включающего эвристические и формальные процедуры обнаружения субъектных столбцов, определения предикатных связей, нормализации единиц измерения и временных интервалов, а также алгоритмы построения и валидации RDF-триплетов. В разделе рассматриваются критерии качества семантической интерпретации, показатели согласованности разметки и методы разрешения неоднозначностей, возникающих при обработке табличных данных различной степени структурированности.

В пункте 3.4 вводятся SWRL-правила, усиливающие семантический уровень модели и обеспечивающие автоматическое присвоение классов качества воды на основании нормативных порогов модуля WR_KZ, связывание наблюдений с регионами и временными интервалами, а также индукцию новых логических связей. Дополнительно анализируется практическая реализация этих правил в среде GraphDB, шаблоны SPARQL-запросов и примеры пространственно-временной геовизуализации результатов.

Научная новизна раздела заключается в разработке многоуровневой архитектуры семантической обработки гидрологических данных, интегрирующей процессы интерпретации таблиц, онтологического выравнивания, логического вывода и геопространственной визуализации в рамках единого воспроизводимого конвейера. Предложена формализованная схема интеграции международных гидрологических онтологий с национальными нормативными и социально-географическими модулями, обеспечивающая расширяемость модели и её адаптацию к региональному контексту. Разработан итеративный метод семантической интерпретации таблиц, обеспечивающий автоматизированное преобразование гетерогенных табличных данных в онтологически согласованный граф знаний с сохранением контекстной информации. Внедрение SWRL-правил позволяет перейти от статической интеграции данных к динамической модели логического вывода, способной автоматически выявлять нормативные несоответствия и формировать новые знания. Тем самым создаётся теоретически и технологически обоснованная основа для построения интеллектуальной геоинформационной системы мониторинга водных ресурсов.

3.1 Многоуровневая архитектура онтологически-ориентированной геоинформационной модели для извлечения знаний

Целью настоящего подраздела является формализация и теоретическое обоснование многоуровневой архитектуры геоинформационной модели, основанной на принципах семантической интерпретации данных,

онтологического выравнивания и логического вывода. Предлагаемая архитектура определяет структурный каркас методологии, обеспечивающий переход от неструктурированных и полуформализованных табличных источников к воспроизводимому графу знаний, пригодному для аналитической обработки и пространственно-временной визуализации.

Рисунок 3.1 иллюстрирует целостный конвейер «таблица → RDF → онтология → запросы/гео-визуализация», на котором основана предлагаемая методология.

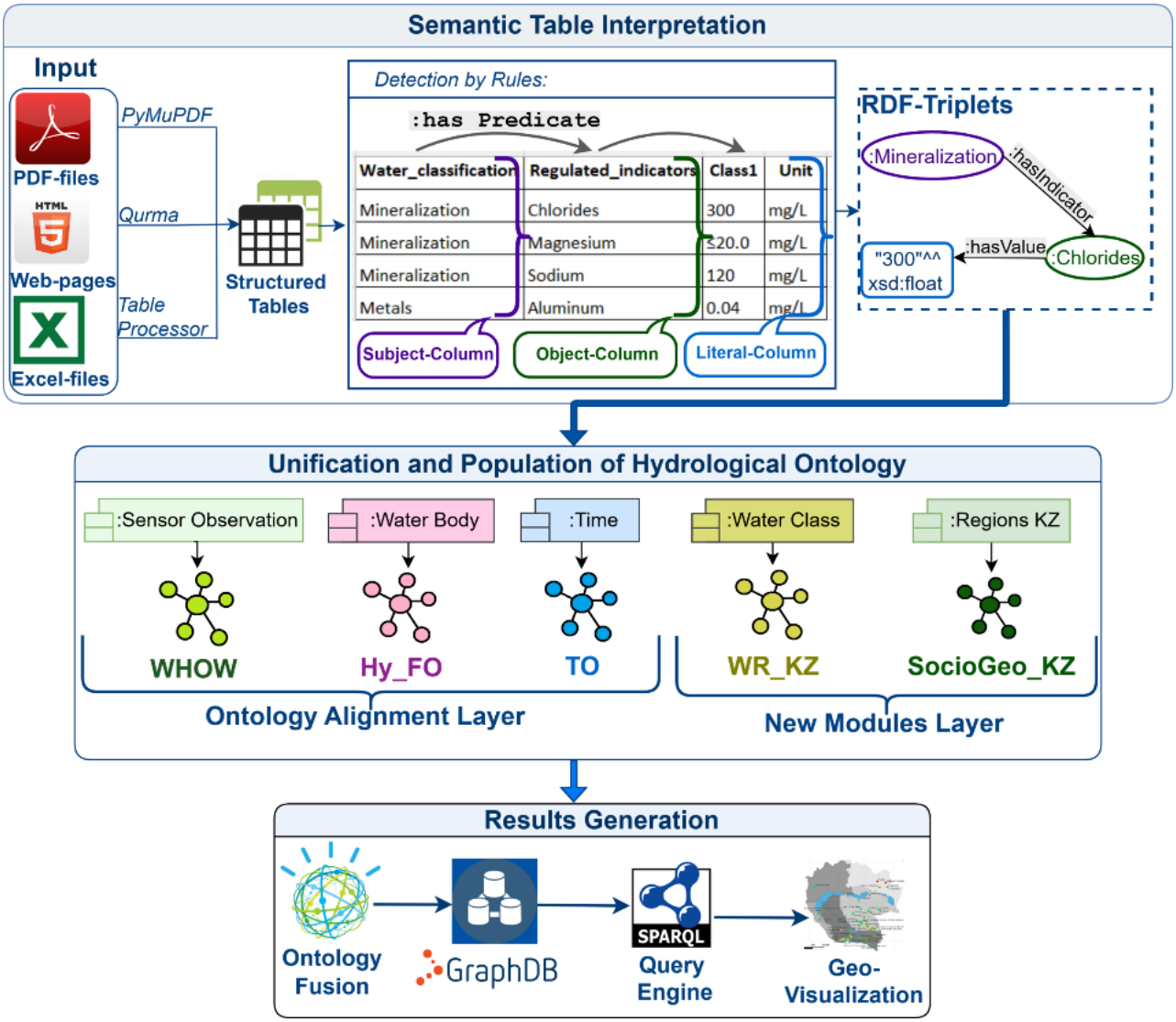


Рисунок 3.1 – Многоуровневый конвейер архитектуры геоинформационной модели на основе онтологии

Архитектура построена модульно и охватывает три логических уровня:

(i) семантическая интерпретация табличных данных (STI), где для разных форматов данных использовались разные методы, для более точного извлечения;

(ii) унификация и пополнение гидроонтологии, где извлченнные триплеты либо добавляются в существующие международные

онтологий, если есть соответствующие субъекты, либо пополняют новые модули, которые характерны только для водных ресурсов Казахстана;

(iii) генерация результатов (хранилище, запросы, геовизуализация), где пополненные граф знаний применяется для построения гео-информационной модели.

Такое многоуровневое построение обеспечивает декомпозицию задач обработки данных, разделение ответственности между слоями и формирование масштабируемой онтологически-ориентированной системы мониторинга водных ресурсов.

Архитектурный подход направлен на достижение семантической согласованности (semantic interoperability), поддержание онтологического выравнивания (ontology alignment) с международными стандартами и обеспечение воспроизводимости результатов анализа в рамках единого графа знаний.

На первом уровне архитектуры – Semantic Table Interpretation (STI) – реализуется процедура формализованного извлечения, очистки и структурирования информации из гетерогенных источников данных, включая PDF-документы, веб-страницы, отчётные формы и электронные таблицы. Особенность данного этапа заключается в необходимости обработки полуструктурированных и слабо формализованных представлений, в которых табличные данные часто сопровождаются текстовыми комментариями, сложной верстkой, объединёнными ячейками и неоднозначной семантикой заголовков. Для автоматизации процесса извлечения применяются инструменты парсинга и реконструкции табличных структур, такие как PyMuPDF [156], QURMA [159] и TableProcessor [163], обеспечивающие детектирование границ таблиц, восстановление логической структуры строк и столбцов, а также извлечение метаданных, связанных с источником публикации и временным контекстом наблюдений.

Цель данного этапа состоит в приведении всех извлечённых данных к унифицированному формату структурированных таблиц с сохранением семантически значимых атрибутов, включая заголовки, подписи, единицы измерения, ссылки на первоисточник и временные интервалы наблюдений. Сохранение контекстной информации рассматривается как обязательное условие последующего онтологического выравнивания, поскольку именно метаданные обеспечивают корректное сопоставление столбцов таблиц с классами и свойствами доменной онтологии. Потеря контекста на данном этапе приводит к семантической неоднозначности и снижению качества интеграции данных в граф знаний.

В процессе STI реализуется итеративный эвристический механизм семантической разметки таблиц, включающий анализ заголовков, лексико-синтаксические шаблоны, частотные характеристики значений и сопоставление с онтологическим словарём предметной области. На основе совокупности эвристических правил автоматически идентифицируются субъектные (entity

columns), объектные и атрибутивные колонки, устанавливаются предикатные отношения между сущностями и числовыми показателями, а также выполняется нормализация единиц измерения и диапазонов значений с приведением к каноническому формату. Дополнительно осуществляется разрешение неоднозначностей, связанных с омонимией терминов и вариативностью обозначений показателей качества воды.

Результатом функционирования STI-уровня является генерация формально корректных RDF-триплетов вида:

:Subject <predicate> :Object и

:Subject <dataProperty> "value"^^xsd:*,

которые дополняются аннотациями происхождения данных (provenance), временными метками наблюдений и идентификаторами источников. Тем самым обеспечивается переход от табличных представлений к онтологически согласованному семантическому графу, служащему основой для последующей интеграции, логического вывода и пространственно-временного анализа в рамках разработанной геоинформационной модели.

Второй уровень архитектуры – Unification and Population of Hydrological Ontology – выполняет ключевую функцию семантического согласования и концептуальной интеграции данных, полученных на этапе STI, с формализованными доменными моделями. Данный уровень обеспечивает переход от набора разрозненных RDF-триплетов к структурированному онтологическому графу знаний, построенному на основе международно признанных стандартов и дополненному национальными модулями. Центральной задачей этого этапа является реализация процедур онтологического выравнивания (ontology alignment) и семантического сопоставления сущностей, свойств и отношений с использованием формальных механизмов OWL и RDFS.

На данном этапе осуществляется выравнивание с тремя базовыми онтологическими моделями: WHOW/SSN/SOSA [10,42,53], HY_Features [33] и Time Ontology [57], каждая из которых отвечает за отдельное измерение представления гидрологических данных. Онтология WHOW/SSN/SOSA формирует измерительное измерение модели, обеспечивая стандартизированное описание сенсорных наблюдений и их результатов. В рамках данной модели класс sosa:Observation связывается с измеряемым параметром через sosa:observedProperty и с полученным значением через sosa:hasResult, что позволяет унифицировать данные, поступающие из различных сенсорных систем и гидропостов. Дополнительные свойства, описывающие процедуру измерения, используемый сенсор и условия наблюдения, обеспечивают сохранение контекста происхождения данных (provenance), что критически важно для интерпретации и воспроизводимости результатов.

Пространственное измерение модели формируется на основе онтологии HY_Features (Hydrologic Feature Model), которая задаёт формальное описание гидрографических объектов и их иерархических и топологических связей. Классы hyfo:WaterBody, hyfo:River, hyfo:Lake и hyfo:Catchment обеспечивают

типизацию водных объектов, тогда как отношения hyfo:locatedInRegion, hyfo:partOfHydroNetwork и sosa:hasFeatureOfInterest позволяют установить пространственную принадлежность объектов и связать их с соответствующими наблюдениями. Использование данной онтологии создаёт основу геопространственной семантики и обеспечивает совместимость модели с геоинформационными системами и стандартами пространственных данных.

Временное измерение данных реализуется посредством интеграции Time Ontology (OWL-Time), которая обеспечивает формализацию временных характеристик наблюдений. Классы time:Instant и time:Interval позволяют моделировать как точечные измерения, так и продолжительные процессы, а свойства time:hasBeginning, time:hasEnd и time:hasTime задают формальные связи между событиями и временными метками. Такая формализация позволяет анализировать динамику изменений показателей качества воды, выявлять сезонные и межгодовые закономерности, а также обеспечивать корректную агрегацию данных по временным интервалам.

Совместное использование указанных онтологий формирует трёхмерную семантическую модель гидрологических данных, охватывающую измерительный, пространственный и временной аспекты. Для адаптации международных моделей к специфике национальной системы мониторинга параллельно разрабатываются два доменных модуля: Water_Regulations_KZ (WR_KZ), отражающий нормативные пороговые значения и классы качества воды в соответствии с законодательством Республики Казахстан, и SocioGeo_KZ, описывающий социально-географические характеристики регионов, включая административное деление, демографические показатели и хозяйственную нагрузку на водные объекты. Интеграция международных и национальных компонентов осуществляется посредством формальных связей выравнивания (owl:equivalentClass, rdfs:subClassOf, owl:ObjectProperty), что обеспечивает логическую согласованность, расширяемость и возможность повторного использования разработанной модели.

Таким образом, уровень Unification and Population of Hydrological Ontology формирует концептуальную и формально-логическую основу семантической согласованности всей системы. Он объединяет международные стандарты и национальные онтологические расширения в единый когерентный граф знаний, обеспечивающий целостное представление данных о водных ресурсах в пространственно-временном и нормативном контексте, а также создаёт предпосылки для последующего логического вывода и интеллектуального анализа.

Третий уровень архитектуры – Results Generation – представляет собой заключительный и интеграционный этап предложенной методологии, на котором осуществляется переход от семантически согласованного онтологического слоя к аналитически эксплуатируемой системе знаний. Данный уровень обеспечивает объединение всех онтологических компонентов, централизованное хранение данных, выполнение логически обоснованных запросов и последующую пространственно-временную интерпретацию

результатов. Его ключевая функция заключается в трансформации сформированного графа знаний из статической структуры представления данных в динамический инструмент интеллектуального анализа.

На данном этапе реализуется процедура Ontology Fusion, предполагающая слияние международных, национальных и предметно-ориентированных онтологических модулей в единое когерентное семантическое пространство. В ходе слияния обеспечивается устранение дублирующих классов, согласование пространств имён, унификация иерархий и синхронизация свойств на основе формальных механизмов OWL. Итоговый граф знаний загружается в онтологическое хранилище GraphDB [163], выступающее в роли централизованной платформы управления семантическими данными. Выбор GraphDB обусловлен поддержкой механизмов версионирования, инкрементального обновления графа, транзакционной целостности и встроенных средств логического вывода.

Логический вывод реализуется на основе правил RDFS/OWL [46] и расширенных правил SWRL [45], что обеспечивает автоматическую классификацию сущностей, наследование свойств, выявление скрытых зависимостей и индукцию новых семантических связей. В частности, применение SWRL-правил позволяет автоматически сопоставлять измеренные показатели с нормативными порогами, присваивать класс качества воды и выявлять пространственно-временные закономерности без дополнительной ручной обработки данных. Тем самым граф знаний приобретает свойства самообогащающейся структуры, способной к динамическому расширению и адаптации при поступлении новых данных.

Язык SPARQL (SPARQL Protocol and RDF Query Language) используется в качестве основного инструмента выборки, агрегации и аналитической обработки информации в онтологическом хранилище. Его декларативная природа позволяет формулировать сложные исследовательские запросы, включающие множественные соединения, фильтрацию по временным интервалам, пространственную агрегацию и логические условия. Реализация различных типов запросов – SELECT для получения агрегированных показателей по гидропостам, рекам и временным периодам; CONSTRUCT для построения тематических подграфов и картографических слоёв; ASK и DESCRIBE для проверки целостности данных и навигации по онтологической структуре – обеспечивает гибкость аналитического взаимодействия с графом знаний. Существенным преимуществом данной организации является воспроизводимость результатов: каждый аналитический вывод может быть реконструирован посредством формализованного запроса, что соответствует требованиям научной прозрачности и проверяемости экологических исследований.

Результаты выполнения запросов интегрируются с картографическим модулем, обеспечивающим геопространственную визуализацию данных. На данном этапе осуществляется отображение гидропостов в виде точечных объектов, рек – в виде линейных геометрий, а региональных характеристик – в

форме тематических слоёв. Пространственная визуализация опирается на семантически обогащённые данные графа знаний, что позволяет не только отображать числовые показатели, но и учитывать их нормативную классификацию, временную динамику и взаимосвязь с социально-географическими параметрами регионов. Такая интеграция обеспечивает возможность выявления пространственно-временных закономерностей, анализа сезонных колебаний, оценки соответствия санитарным стандартам и сопоставления экологических показателей с факторами антропогенной нагрузки.

Таким образом, уровень Results Generation завершает полный цикл семантической обработки данных, обеспечивая преобразование онтологически интегрированного графа знаний в инструмент интеллектуального геоинформационного анализа. Он объединяет механизмы логического вывода, онтологического поиска, воспроизводимых запросов и пространственно-временной визуализации, формируя технологически и концептуально завершённую систему мониторинга состояния водных ресурсов.

3.2 Модульная архитектура гидроонтологии: выравнивание с международными онтологиями и расширение национальными модулями

Современные задачи мониторинга водных ресурсов требуют не только накопления данных, но и их формального семантического представления, обеспечивающего интероперабельность, логическую согласованность и возможность автоматизированного анализа. В условиях интеграции разнородных источников – гидрохимических измерений, нормативных документов, геопространственных характеристик и социально-экономических показателей – особое значение приобретает построение онтологической модели, способной объединить различные измерения данных в рамках единого концептуального пространства.

На рисунке 3.2 представлена структура разработанной гидроонтологии, отражающая её модульную организацию и распределение функциональных компонентов. Архитектура модели основана на принципе разделения на два взаимосвязанных слоя: слой онтологического выравнивания (Ontology Alignment Layer), обеспечивающий повторное использование и согласование международных стандартов, и слой национальных расширений (New Modules Layer), адаптирующий модель к региональному контексту мониторинга водных ресурсов Казахстана. Такая композиция позволяет обеспечить баланс между универсальностью и предметной спецификой.

Разработанная онтологическая модель направлена на формирование единого семантического пространства, объединяющего данные наблюдений за водными ресурсами, нормативы качества воды и социально-географический контекст [165]. В рамках модели выделяются несколько взаимодополняющих доменных блоков. Международный слой включает онтологии SSN/SOSA для описания сенсорных наблюдений и их результатов, HY_Features для формализации гидрографических объектов (реки, озёра, водосборные бассейны, гидросети) и OWL-Time для представления временных интервалов и моментов

измерений

Национальный слой представлен модулями WR_KZ и SocioGeo_KZ. Модуль WR_KZ формализует систему нормативов качества воды, включая классы водных объектов, регулируемые показатели, гигиенические стандарты, единицы измерения и предельные значения концентраций. В свою очередь, SocioGeo_KZ описывает социально-географические характеристики регионов, включая координаты, численность населения, показатели рождаемости и смертности, индустриальные предприятия и сельскохозяйственные территории. Включение данного слоя позволяет проводить комплексный анализ взаимосвязи экологических и социально-экономических факторов.

Таким образом, модульная архитектура гидроонтологии обеспечивает структурную декомпозицию предметной области на взаимосвязанные семантические компоненты, объединённые через механизмы выравнивания (owl:equivalentClass, rdfs:subClassOf, объектные свойства сопоставления). В дальнейшем подробно рассматриваются принципы выравнивания международных онтологий и концептуальное обоснование национальных модулей, представленных на рисунке 3.2.

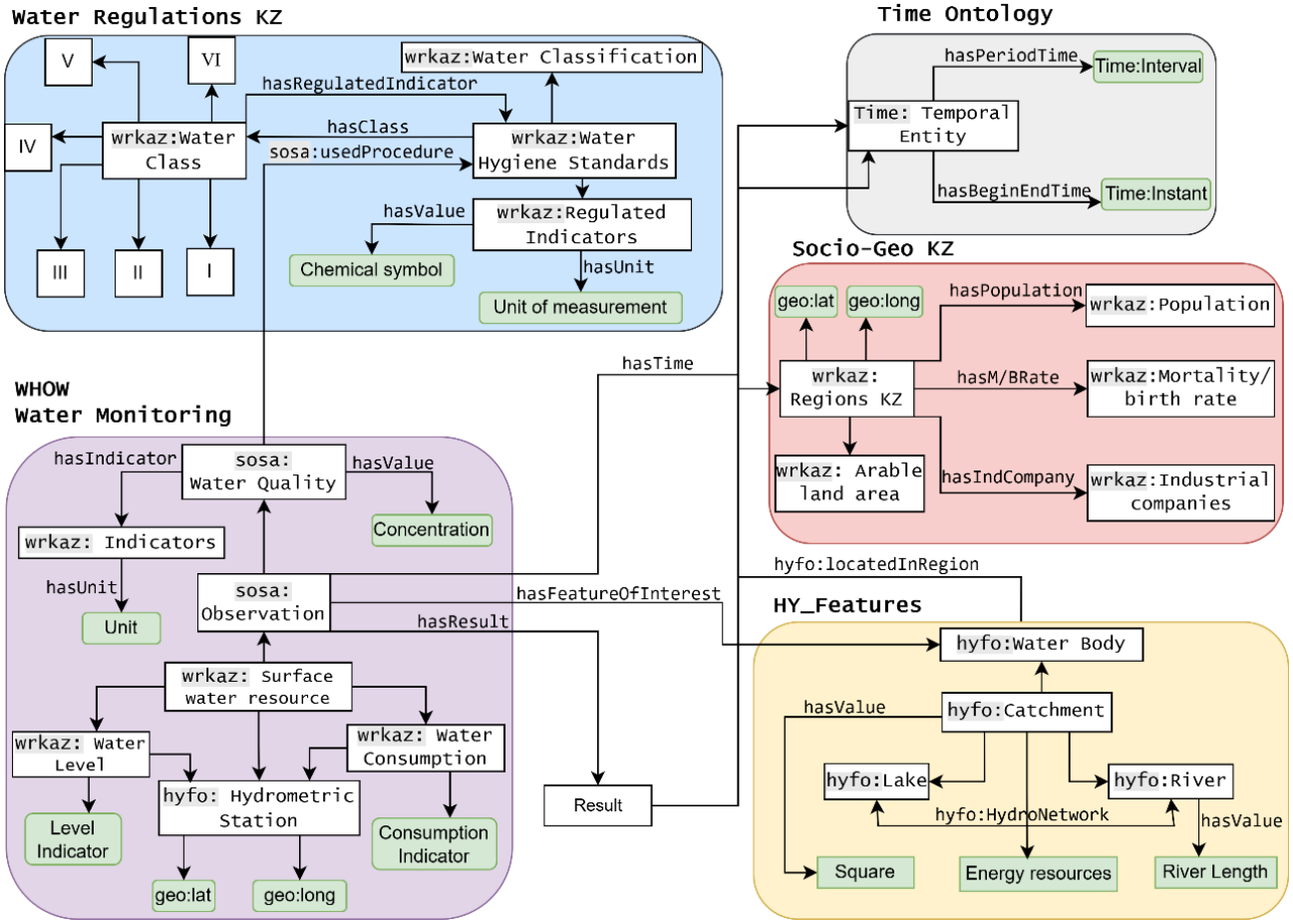


Рисунок 3.2 - Выравнивание гидро-онтологической модели с международными онтологиями и её расширение модулями «Нормативы качества воды РК» и «Социально-географический слой»

Проектирование модели осуществлялось на основе трёх методологических

принципов.

Во-первых, реализовано повторное использование терминов: при совпадении с международными онтологиями применяются отношения *owl:equivalentClass/Property*, а при частичном соответствии – *rdfs:subClassOf* и *rdfs:subPropertyOf*.

Во-вторых, обеспечено чёткое разграничение уровней: фундаментальные концепции наблюдений, времени и гидрографической сети описаны в международных онтологиях, тогда как национальные особенности – классы качества воды, пороговые значения, социально-экономические показатели – представлены в пространствах имён *wrkaz:* и *socio:*.

В-третьих, реализована ссылочная совместимость: каждому новому классу и свойству присваивается устойчивый URI (например, *wrkaz:* для нормативных параметров и *socio:* для социально-географического слоя), что гарантирует масштабируемость и интероперабельность всей системы.

По сравнению со стандартными онтологиями предложенная гидроонтологическая модель обладает большей предметной спецификой, адаптированной к региональному контексту Казахстана. Она сочетает универсальные принципы международных стандартов и локальные расширения, обеспечивая как глобальную совместимость, так и национальную актуальность.

3.2.1 Сравнение стандартных онтологий и разработанной гидроонтологической модели

Современные международные онтологии, такие как SOSA/SSN, HY_Features и OWL-Time, представляют собой основу для описания наблюдений, пространственных объектов и временных отношений в рамках концепции Semantic Web. Однако их универсальный характер ограничивает возможность детального учёта национальных, экологических и правовых особенностей. Например, SOSA/SSN описывает акты наблюдений и результаты измерений, но не включает механизмы оценки данных с учётом нормативов качества воды; HY_Features формализует пространственную структуру гидрологических объектов, не связывая их с химико-физическими параметрами загрязнения; OWL-Time обеспечивает описание временных интервалов и событий, но не поддерживает интеграцию временных характеристик с экологическими или социальными факторами.

В отличие от этих стандартных онтологий, разработанная гидроонтологическая модель ориентирована на комплексное представление данных мониторинга водных ресурсов Республики Казахстан. Она объединяет измерительные, нормативные, пространственно-временные и социально-географические аспекты в едином графе знаний, обеспечивая не только структурное описание данных, но и их семантическую интерпретацию.

Ключевые преимущества гидроонтологической модели заключаются в следующем:

1. Семантическая совместимость – базовые элементы модели выровнены с SOSA/SSN, HY_Features и OWL-Time через механизмы *rdfs:subClassOf*,

owl:equivalentClass и *rdfs:subPropertyOf*, что обеспечивает интероперабельность и переносимость данных;

2. Нормативное расширение – модуль *WR_KZ* формализует классы качества воды, предельные концентрации и национальные пороговые значения в соответствии с экологическим законодательством РК;

3. Социально-географическая интеграция – модуль *SocioGeo_KZ* связывает гидрологические показатели с демографическими и промышленными данными, создавая возможность для междисциплинарного анализа;

4. Аналитическая направленность – использование SWRL-правил и SPARQL UPDATE обеспечивает автоматический логический вывод, формирование новых знаний и поддержку интеллектуальных запросов.

Таким образом, разработанная гидроонтологическая модель не заменяет существующие международные стандарты, а расширяет их функциональность, адаптируя общие концепции к национальной системе водного мониторинга. Она обеспечивает переход от статического хранения данных к динамической семантической интерпретации, что позволяет не только анализировать текущее состояние водных ресурсов, но и прогнозировать их изменения с учётом нормативных и социальных факторов.

3.2.2 Модуль наблюдений и качества воды (SOSA/SSN/WHOW)

Модуль Water Monitoring формирует измерительное ядро разработанной гидроонтологии и основан на спецификациях W3C WHOW/SOSA/SSN [10, 42, 53], предназначенных для формализации сенсорных наблюдений и их результатов. Выбор данной группы онтологий обусловлен их международным признанием, совместимостью с инфраструктурами открытых данных и возможностью интеграции с распределёнными системами мониторинга. Использование стандартов W3C обеспечивает семантическую интероперабельность модели и её согласование с современными практиками публикации экологических данных в формате Linked Open Data.

В рамках модуля центральным элементом является класс sosa:Observation, моделирующий акт измерения определённого параметра в конкретном пространственно-временном контексте. Наблюдение связывается с измеряемым свойством посредством отношения sosa:observedProperty, которое в разработанной модели реализуется через класс wrkz:Indicators, представляющий перечень контролируемых гидрохимических показателей (например, «Хлориды», «Сульфаты», «Железо (общее)», «Минерализация»). Результат измерения фиксируется через свойство sosa:hasResult, обеспечивающее связь между наблюдением и числовым значением концентрации или уровня показателя. Дополнительно учитываются атрибуты, описывающие процедуру измерения (sosa:usedProcedure), измерительное устройство и объект наблюдения (sosa:hasFeatureOfInterest), что позволяет сохранить контекст получения данных и обеспечить их воспроизводимость.

Особое внимание уделяется нормализации единиц измерения и обеспечению сопоставимости количественных показателей. Для этого

интегрируются онтологии QUDT (Quantities, Units, Dimensions and Data Types) и OM (Ontology of Units of Measure), что позволяет формально описывать размерности, единицы и типы данных. Свойство hasUnit связывает измеренное значение с соответствующей единицей измерения, обеспечивая корректность арифметических операций, агрегации и сравнительного анализа. Такая нормализация исключает неоднозначности, возникающие при использовании различных форматов представления концентраций (мг/л, г/м³ и др.), и способствует унификации данных, поступающих из разных источников. Онтология SOSA (Sensor, Observation, Sample, and Actuator) рассматривает наблюдение как центральную сущность, описывающую факт измерения определённого свойства объекта в конкретный момент времени. Связи между наблюдением, измеряемым параметром и результатом измерения формализуются посредством свойств sosa:observedProperty, sosa:hasResult и sosa:hasFeatureOfInterest. Расширение SSN (Semantic Sensor Network) дополняет данную модель описанием сенсоров, процедур измерения и условий наблюдения, обеспечивая полноту контекстной информации. В рамках проекта WHOW (Water Health Open Knowledge) данная структура была адаптирована для задач мониторинга качества воды.

Официальная модель онтологии WHOW/SOSA/SSN, иллюстрирующая структуру сенсорных наблюдений и их связей, представлена на рисунке 3.3 и служит концептуальной основой для проектирования данного модуля.

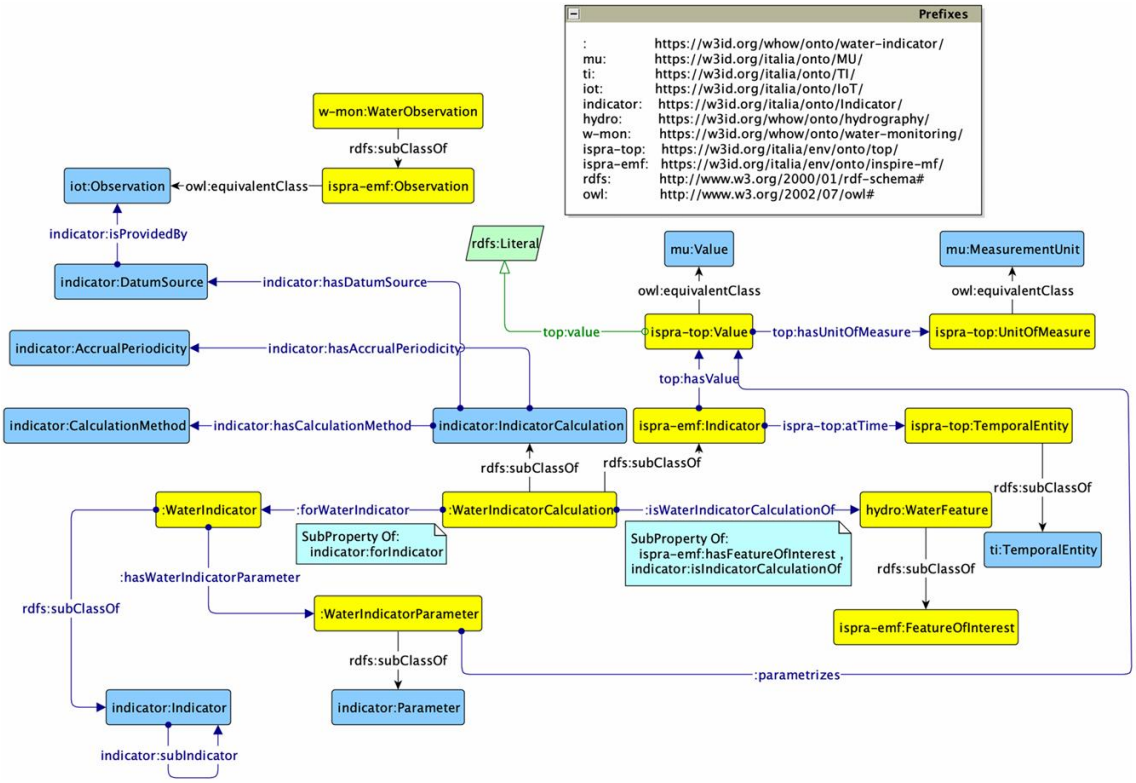


Рисунок 3.3 – Официальная Модель онтологии WHOW/SOSA/SSN [53]

Представленная модель послужила концептуальной основой для

разработки авторской версии модуля Water Monitoring, ориентированной на специфику гидрологического мониторинга в Республике Казахстан (Рис.3.4). В разработанной модели сохранён базовый измерительный паттерн SOSA, однако он дополнен специализированными доменными классами и свойствами, отражающими структуру национальной системы наблюдений.

В структуре разработанной модели классы wrkz:WaterQuality и wrkz:Indicators определены как специализированные подмножества наблюдаемых свойств и выровнены с концепциями SOSA посредством отношения rdfs:subClassOf. Это позволяет сохранить семантическую совместимость с международным стандартом, одновременно расширяя модель национально-специфичными показателями и классификациями.

Интеграция модуля наблюдений с пространственным и временным слоями онтологии обеспечивает возможность сопоставления данных по гидропостам, рекам, бассейнам и временным интервалам наблюдений. Благодаря этому создаются предпосылки для построения временных рядов, анализа сезонной динамики показателей качества воды и выявления устойчивых экологических тенденций.

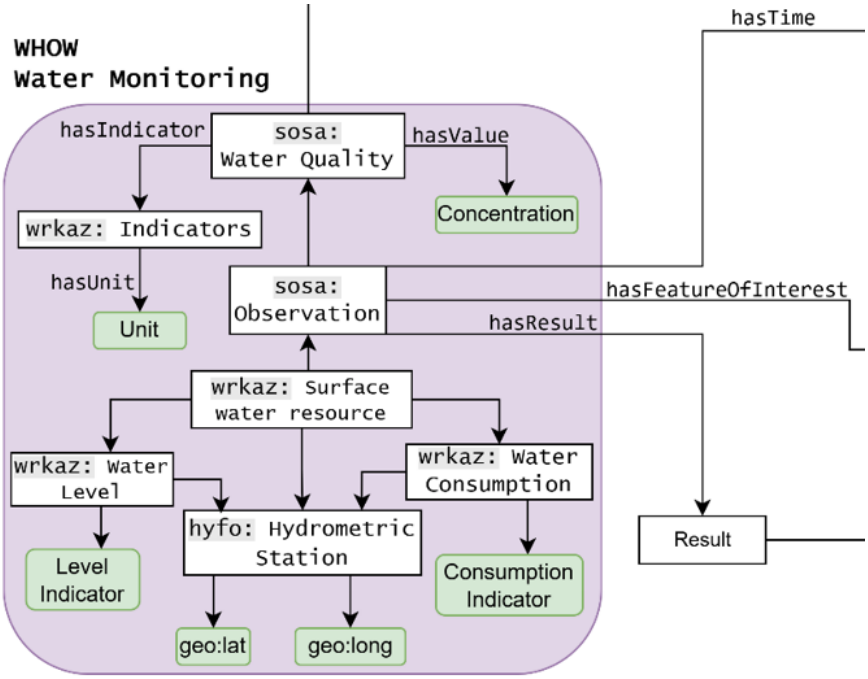


Рисунок 3.4 – Модуль наблюдений и качества воды (SOSA/SSN / WHOW)

3.2.3 Модуль гидрографических объектов (HY_Features)

Модель HY_Features [33], разработанная OGC, используется для описания пространственных характеристик гидрографической сети. Абстрактный класс *hyfo:WaterBody* объединяет основные типы водных объектов: *hyfo:River*, *hyfo:Lake*, *hyfo:Catchment* и *hyfo:HydroNetwork*. Морфометрические и пространственные параметры представлены свойствами *hyfo:RiverLength*, *hyfo:Square* и *hyfo:EnergyResources*. Связь водных объектов с наблюдениями осуществляется через *sosa:hasFeatureOfInterest*, а координаты гидропостов фиксируются свойствами *geo:lat* и *geo:long*. Этот модуль является

пространственным ядром модели, обеспечивая географическую основу для интеграции с ГИС-системами и пространственного анализа.

Дополнительно следует отметить, что использование модели HY_Features в сочетании с расширениями HyFO и стандартами SSN/SOSA обеспечивает формально согласованное представление как геометрических, так и семантических аспектов водных объектов, что существенно повышает уровень интероперабельности создаваемой онтологической модели. Применение унифицированных классов и свойств позволяет обеспечить корректное сопоставление гидрографических сущностей при интеграции данных из разнородных источников, включая гидрометеорологические наблюдения, статистические отчёты и геопространственные базы. геоинформационными сервисами. На Рисунках 3.5 и 3.6 приведены модули онтологии HY_Features и какие классы из данной онтологии были импортированы в нашу гидро-онтологию для Казахстана.

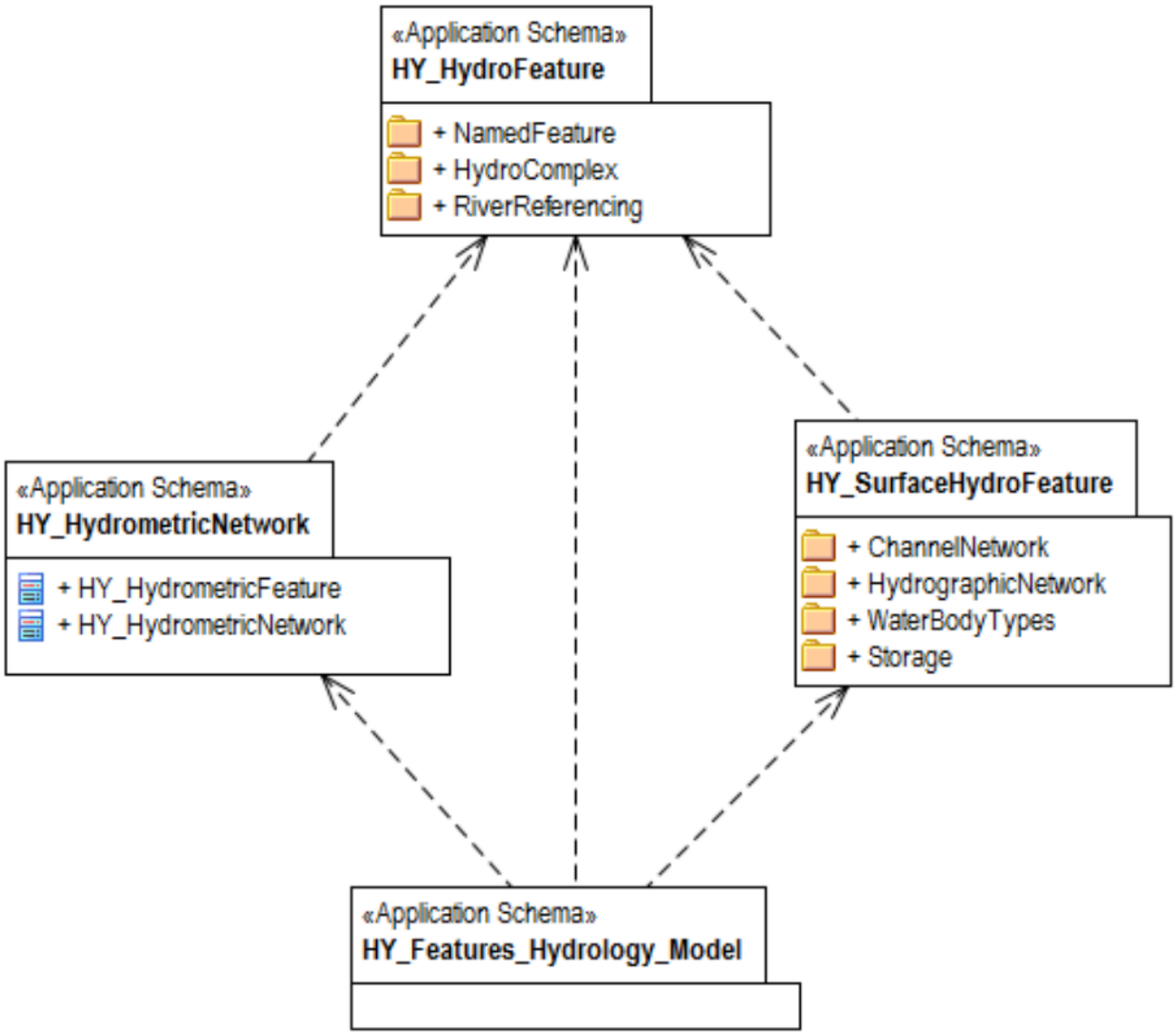


Рисунок 3.5 – Официальная модель онтологии HY_Features [33]

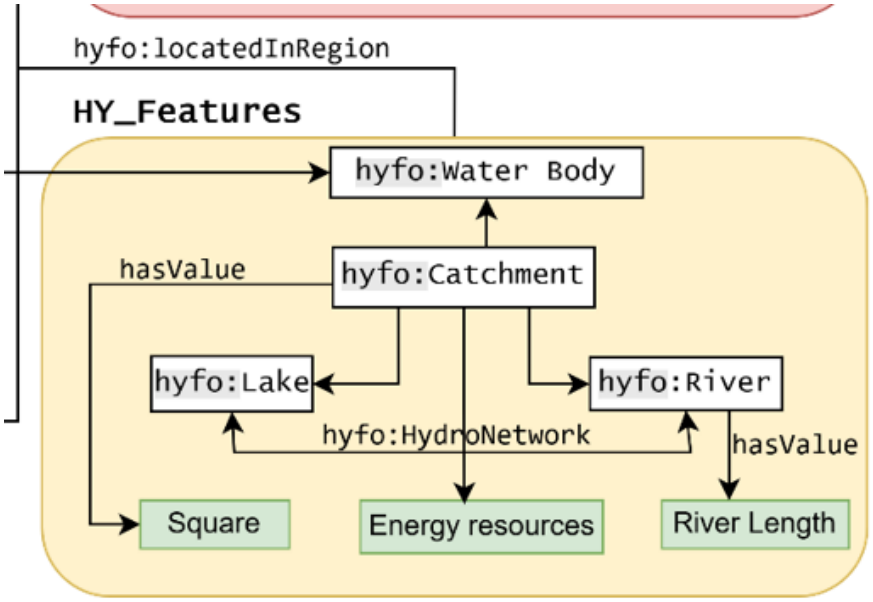


Рисунок 3.6 - Модуль гидро-онтологии для Казахстана с импортированными классами из онтологии HY_Features

Временная привязка OWL-Time модели построена на онтологии *W3C OWL-Time* [57]. Класс *time:Instant* описывает отдельные временные метки наблюдений, а *time:Interval* – периоды мониторинга. Свойства *time:hasBeginning*, *time:hasEnd* и *time:hasTime* используются для аннотирования наблюдений и временной агрегации показателей. Введение временных связей позволяет выполнять анализ динамики, выявлять сезонные закономерности и сопоставлять показатели по различным периодам наблюдения. На Рис.3.7 приведена модуль временной онтологии, которая была импортирована в нашу онтологию.

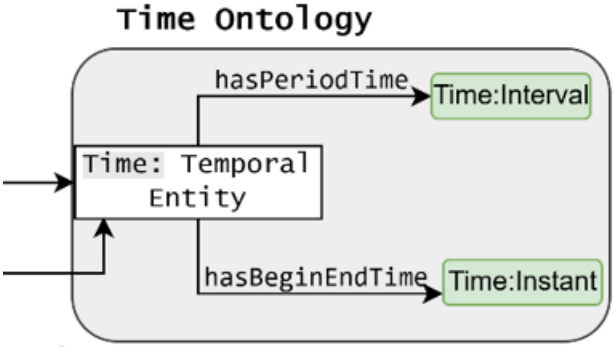


Рисунок 3.7 – Модуль временной онтологии OWL-Time

3.2.4 Национальный модуль нормативов качества воды – WR_KZ

Национальный модуль WR_KZ (Water Regulations Kazakhstan) разработан с целью формализации действующей нормативно-правовой базы Республики Казахстан в области оценки качества поверхностных вод и её интеграции в единую онтологическую структуру. В отличие от международных онтологий, ориентированных преимущественно на универсальные измерительные и пространственные паттерны, данный модуль отражает специфику национальных

санитарно-гигиенических требований, классификационных шкал и предельно допустимых концентраций, установленных нормативными актами.

Модуль wrkaz: обеспечивает семантическое представление нормативных правил, регламентирующих оценку состояния водных объектов, включая гигиенические стандарты, перечень регулируемых показателей и пороговые значения концентраций. Класс wrkz:WaterHygieneStandards описывает нормативные документы и методики проведения измерений, включая ссылку на источник правового акта, тип процедуры и область применения стандарта. Свойство sosa:usedProcedure используется для согласования измерительного процесса с соответствующей нормативной методикой, что позволяет обеспечить корректность сопоставления результатов наблюдений с установленными требованиями.

Класс wrkz:RegulatedIndicators формализует перечень контролируемых показателей качества воды, включая химические элементы, соединения и интегральные параметры (например, минерализация, БПК, содержание тяжёлых металлов). Для каждого показателя задаются единицы измерения, выровненные с онтологиями QUDT и OM, а также пороговые значения через свойство wrkz:hasThreshold, позволяющее хранить нормативные границы в машиночитаемом формате. Это обеспечивает возможность автоматизированного сравнения фактических концентраций с установленными пределами.

Класс wrkz:WaterClass отражает национальную систему классификации качества воды по степеням загрязнения (I–VI классы), принятую в нормативных документах Республики Казахстан. Свойство wrkz:definesWaterClass связывает конкретные пороговые значения с соответствующим классом воды, формируя формально определённую логическую структуру классификации. Такая модель позволяет реализовать механизм автоматической категоризации состояния водного объекта на основе совокупности измеренных показателей.

Особое значение имеет механизм семантического выравнивания модуля WR_KZ с международными онтологиями. Классы и свойства национального уровня интегрируются в общую модель посредством отношений rdfs:subClassOf, owl:equivalentClass и сопоставления процедур измерений через sosa:usedProcedure. Это обеспечивает достижение семантической интероперабельности и позволяет использовать единый механизм логического вывода для международных и национальных компонентов модели. В частности, применение SWRL-правил (см. гл. 3.4) позволяет автоматически сопоставлять результаты наблюдений с нормативными порогами и присваивать класс качества воды без ручного анализа.

Таким образом, модуль WR_KZ выполняет функцию нормативного слоя разработанной гидроонтологии, обеспечивая формальную интеграцию правовых требований в структуру графа знаний. Его включение позволяет перейти от описательной модели мониторинга к регуляторно-ориентированной системе анализа, способной выявлять несоответствия санитарным нормам, формировать классификационные выводы и поддерживать принятие управленческих

решений.

Модель нормативных правил по установлению класса воды в соответствии с действующими водными правилами Республики Казахстан, опубликованными на официальном портале zan.kz, представлена на рисунке 3.8. Данный рисунок иллюстрирует логические связи между измеряемыми показателями, пороговыми значениями и классами качества воды, формализованными в рамках модуля WR_KZ.

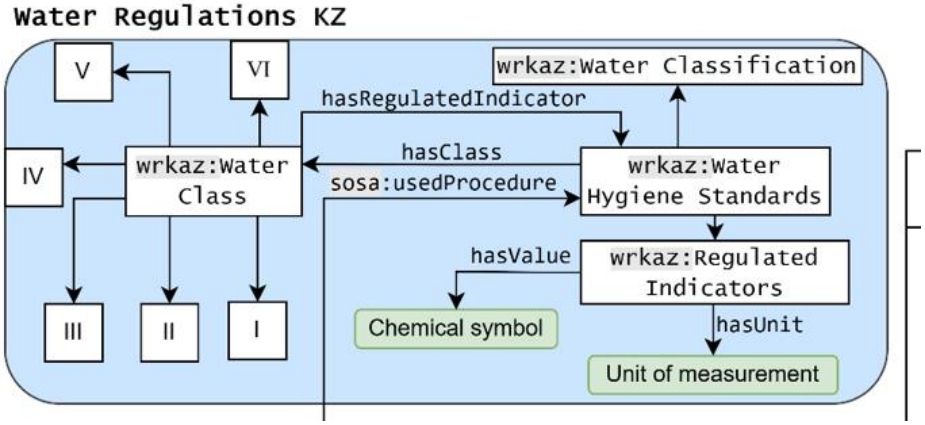


Рисунок 3.8 – Национальный модуль нормативов качества воды

Национальный социально-географический модуль SocioGeo_KZ разработан с целью интеграции гидрологических данных с социально-экономическими характеристиками регионов, формируя междисциплинарный аналитический слой в структуре разработанной гидроонтологии. Включение данного модуля обусловлено необходимостью перехода от изолированного анализа показателей качества воды к комплексной оценке факторов антропогенной нагрузки, демографической динамики и хозяйственной активности, оказывающих влияние на состояние водных ресурсов.

В структуре модуля центральным элементом является класс wrkz:RegionsKZ, описывающий административно-территориальные единицы Республики Казахстан. Для обеспечения пространственной идентификации регионов используются свойства geo:lat и geo:long, позволяющие задать географические координаты и обеспечить совместимость с геоинформационными системами. Такое формальное описание пространственных характеристик создаёт основу для картографической визуализации и пространственного анализа взаимосвязей между экологическими и социально-экономическими показателями.

Социально-экономические параметры представлены рядом специализированных классов, включая wrkz:Population, отражающий численность населения региона; wrkz:MortalityBirthRate, описывающий демографические коэффициенты рождаемости и смертности; wrkz:IndustrialCompanies, характеризующий уровень индустриальной активности; а также wrkz:ArableLandArea, фиксирующий площадь сельскохозяйственных угодий. Эти сущности формируют количественную

основу для оценки уровня антропогенного воздействия на водные объекты.

Связь гидрологических и социально-экономических данных реализуется посредством свойств socio:hasPopulation, socio:hasM/BRate, socio:hasIndCompany, а также пространственного отношения hyfo:locatedInRegion, связывающего водные объекты (реки, озёра, водосборные бассейны) с соответствующими административными единицами. Такая семантическая интеграция позволяет моделировать комплексные зависимости между качеством воды, плотностью населения, промышленной нагрузкой и аграрной деятельностью, обеспечивая основу для межотраслевого анализа.

Включение модуля SocioGeo_KZ расширяет функциональные возможности графа знаний, позволяя формировать запросы, объединяющие экологические и социальные показатели, выявлять корреляции между демографическими характеристиками и изменением гидрохимических параметров, а также оценивать пространственную неоднородность антропогенного воздействия. Таким образом, модуль выполняет роль контекстуального слоя, дополняющего измерительное, пространственное и нормативное измерения модели.

На рисунке 3.9 представлена структура национального социально-географического модуля, демонстрирующая классы и связи, обеспечивающие интеграцию региональных социально-экономических характеристик с гидрологическими объектами в рамках единого онтологического пространства.

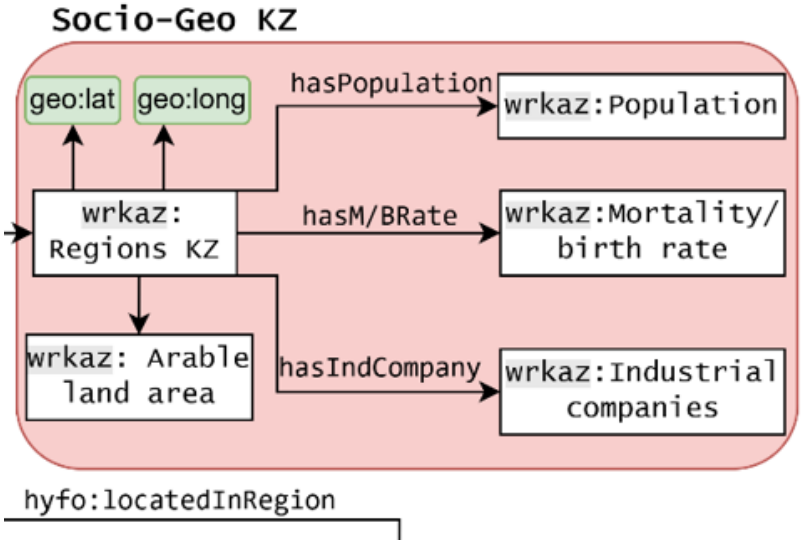


Рисунок 3.9 – Национальный социально-географический модуль – SocioGeo_KZ

Ниже приведены предикаты, через которых осуществляется семантическое связывание (alignment patterns):

1. Наблюдение → Объект → Время: В первом шаблоне наблюдение (*sosa:Observation*) связывается с водным объектом (*hyfo:WaterBody*) и временной меткой (*time:Instant/Interval*) посредством свойств *sosa:hasFeatureOfInterest* и *time:hasTime*.

2. Наблюдение → Норматив: Во втором шаблоне устанавливается связь между наблюдением и нормативными элементами: *wrkaz:WaterHygieneStandards – wrkaz:hasRegulatedIndicator → wrkaz:RegulatedIndicators*, при этом *wrkaz:WaterClass* служит целевой классификацией, обеспечивая интерпретацию результатов наблюдений в рамках нормативных порогов. (гл. 3.3).

3. Объект/Пост → Регион: В третьем шаблоне отражается пространственная привязка объектов к регионам: *hyfo:WaterBody – hyfo:locatedInRegion → wrkaz:RegionsKZ*, где координаты гидропостов задаются свойствами *geo:lat* и *geo:long*.

Эти паттерны формируют основу онтологического вывода и обеспечивают семантическую согласованность данных между различными уровнями модели.

В рамках предложенной модульной архитектуры разработан набор компетентностных запросов, позволяющих проверять корректность модели и демонстрировать её аналитический потенциал:

1. Определение класса воды по посту и дате с учётом порогов WR_KZ: Данный запрос предназначен для автоматического присвоения классов качества воды в соответствии с национальными нормативами модуля *WR_KZ*. Используя данные наблюдений, связываемые через свойства *sosa:Observation*, *time:hasTime* и *wrkaz:hasThreshold*, система позволяет вычислять принадлежность результатов измерений к одному из шести классов загрязнения (I–VI). Такой механизм обеспечивает оперативную классификацию и выявление участков с превышением допустимых концентраций, что важно для государственного экологического надзора и анализа динамики загрязнений.

2. Выборка рек с превышением нормативных показателей и их группировка по регионам**:** Этот запрос реализует перекрёстный анализ данных, используя связи между модулями *HY_Features*, *WR_KZ* и *SocioGeo_KZ*. Он позволяет отобрать водные объекты, где наблюдались превышения по отдельным показателям (*wrkaz:RegulatedIndicators*), и сгруппировать результаты по административным регионам (*wrkaz:RegionsKZ*). Таким образом, обеспечивается возможность пространственной агрегации данных и построения тематических карт, отражающих экологическую ситуацию в разрезе территорий.

3. Сопоставление динамики качества воды с демографическими и экономическими показателями: Третий запрос направлен на выявление корреляционных связей между состоянием водных объектов и социально-экономическими характеристиками регионов. На основе свойств *socio:hasPopulation*, *socio:hasM/BRate* и *socio:hasIndCompany* осуществляется сопоставление изменений показателей качества воды по времени с динамикой населения, уровнем рождаемости и плотностью промышленности. Такой анализ способствует выявлению причинно-следственных зависимостей между экологическим состоянием и антропогенной нагрузкой.

Эти запросы направлены на решение прикладных задач мониторинга и оценки состояния водных ресурсов, а также на установление взаимосвязей между экологическими и социально-экономическими факторами.

Совокупность этих запросов демонстрирует гибкость и интеллектуальную ёмкость разработанной архитектуры. Она позволяет не только переиспользовать международные стандарты для обеспечения базовой совместимости, но и формализовать национальные нормативы и пороговые значения, обеспечивая тем самым воспроизводимое и масштабируемое пополнение графа знаний актуальными данными мониторинга и социально-географической статистики.

3.3 Метод итеративной семантической интерпретации таблиц для пополнения и обновления графов знаний

Современные методы обработки табличных данных играют ключевую роль в задачах автоматизированного формирования, расширения и актуализации графов знаний. В гидрологическом домене данные поступают из разнородных источников – гидрометрических станций, лабораторных анализов, ведомственных отчётов, архивных документов и веб-ресурсов – и, как правило, представлены в неструктурированном или частично структурированном виде. Такие таблицы могут содержать неоднозначные заголовки, объединённые ячейки, неполные единицы измерения и вариативные терминологические обозначения показателей. Для их систематизации и приведения к онтологически совместимому формату в настоящей работе разработан итеративный метод семантической интерпретации таблиц (STI – Semantic Table Interpretation), обеспечивающий последовательное извлечение знаний и их интеграцию в RDF-граф [166].

Основной целью метода является формирование семантически корректных RDF-триплетов, отражающих связи между объектами наблюдений, измеряемыми показателями и количественными результатами, а также их последующее включение в существующую онтологическую модель посредством механизма SPARQL UPDATE. Предложенный подход сочетает эвристические правила, лексико-семантический анализ и контекстную нормализацию, обеспечивая устойчивость к шумным данным, неполным структурам и отсутствию явных идентификаторов. Центральной особенностью алгоритма является реализация итеративного цикла «ячейка-за-ячейкой», в рамках которого каждый новый проход уточняет ранее сформированные связи, повышая точность и полноту интерпретации.

Методологически алгоритм основан на предположении, что табличные структуры содержат скрытую семантику: строки таблицы соответствуют конкретным наблюдениям или сущностям, а столбцы – их атрибутам и отношениям. Таким образом, задача STI сводится к реконструкции неявной семантической схемы таблицы и отображению её структуры в модель «субъект – предикат – объект/литерал».

Процесс семантической интерпретации реализуется поэтапно и включает следующие логические блоки:

1. Определение субъектных столбцов;
2. Идентификация предикатов и формирование кандидатных связей;
3. Типизация и маркировка атрибутивных (литеральных) колонок;

4. Генерация RDF-триплетов;
5. Интеграция и обновление графа знаний.

Общая последовательность работы алгоритма представлена на рисунке 3.10, где отображена диаграмма итеративного процесса, отражающая поток данных от исходной таблицы к формированию RDF-триплетов и их загрузке в граф знаний.

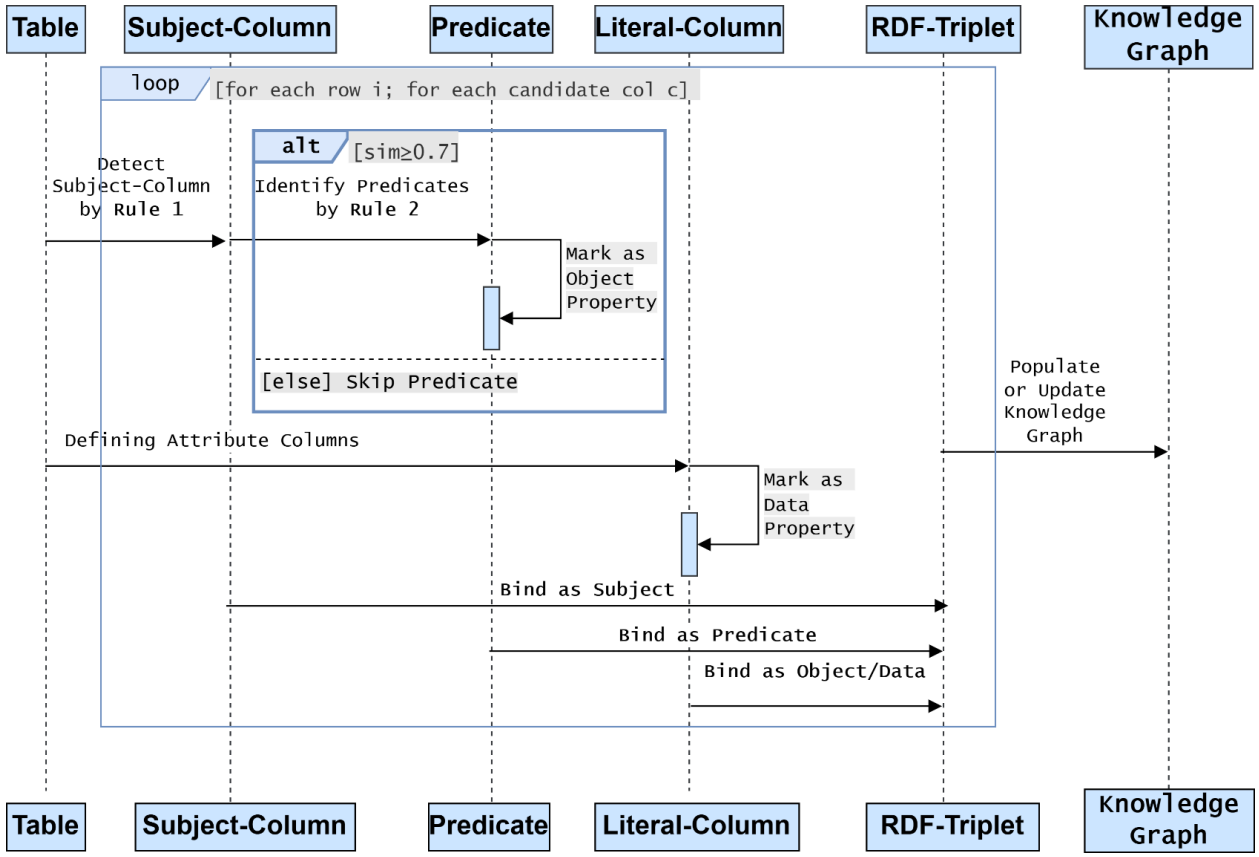


Рисунок 3.10 – Диаграмма последовательности итеративного алгоритма по STI для пополнения и обновления графов знаний

На первом этапе осуществляется контекстная нормализация данных (формула 3.1), включающая приведение строк к унифицированному регистру, удаление служебных символов, стандартизацию единиц измерения и устранение лексических вариаций. Далее производится выявление субъектных столбцов на основе эвристического Правила 1 (Таблица 3.1). Для каждого столбца T_j вычисляется метрика субъектности (формула 3.2), учитывающая уникальность значений, долю текстовых ячеек, частоту повторений и позиционную значимость в структуре таблицы. Столбец с максимальным значением данной метрики определяется как кандидат на роль главной сущности (Subject Column). Это позволяет корректно выделять ключевые столбцы даже при наличии пропусков, дублирований и шумовых данных.

На втором этапе алгоритм осуществляет идентификацию предикатов на основе эвристического Правила 2 (Таблица 3.1). Для каждой пары «субъектная ячейка – потенциальный объект» формируется множество кандидатных

триплетов (формула 3.3).

С целью сопоставления текстовых значений таблицы с элементами онтологии применяется нормализованная метрика Левенштейна (формула 3.4), позволяющая количественно оценить степень строкового сходства. Связь признаётся релевантной при выполнении условия $sim \geq \tau$, где порог τэмпирически выбран в диапазоне 0.7–0.8 (формула 3.5). Такой выбор обусловлен необходимостью обработки вариативных терминов и шумов, возникающих вследствие морфологических различий, орфографических отклонений или ошибок OCR (например, «Chlorides» vs. «Chloride»).

Для повышения достоверности семантической привязки рассчитывается функция уверенности (формула 3.6), учитывающая степень сходства, частотность употребления термина в домене и контекстное соответствие соседних столбцов. Предикат с максимальным коэффициентом доверия выбирается в качестве финального соответствия, что минимизирует вероятность ложных совпадений и повышает точность извлечения.

На третьем этапе осуществляется типизация атрибутивных колонок. Числовые столбцы маркируются как data properties, а текстовые – как object properties, в зависимости от их соответствия классам онтологии. Выполняется нормализация единиц измерения и проверка допустимых диапазонов значений, что обеспечивает согласованность данных с нормативными требованиями и последующим механизмом SWRL-классификации.

После завершения этапов сопоставления формируются RDF-триплеты вида:

$$:Subject \ < predicate > \ :Object$$

Дополнительно каждому триплету присваиваются метаданные происхождения (source, timestamp), что обеспечивает трассируемость и воспроизводимость результатов.

Заключительным этапом является интеграция сформированных триплетов в существующий граф знаний посредством операций SPARQL UPDATE. Алгоритм поддерживает инкрементальное обновление графа, предотвращая дублирование сущностей через механизм проверки уникальных идентификаторов и URI-нормализации. Итеративная реализация позволяет повторно анализировать таблицу при появлении новых данных или уточнении правил, тем самым повышая адаптивность системы.

Таким образом, предложенный метод STI представляет собой формализованный алгоритм преобразования табличных структур в онтологически согласованный граф знаний, обеспечивающий автоматическое пополнение и обновление семантической модели. Его отличительной особенностью является сочетание эвристических правил, метрик текстового сходства и итеративной коррекции связей, что позволяет эффективно обрабатывать неоднородные гидрологические таблицы и формировать устойчивую основу для дальнейшего логического вывода и аналитической обработки данных.

Таблица 3.1 – Формулы эвристических правил итеративной семантической интерпретаций таблиц

Rule 1	1. Контекстная мера нормализации: $cm_{norm}(T_j) = \max_{c_i \in C} cos_sim(T_j, c_i)$ (3.1)	T_j - кандидаты в столбцы-субъекты, C - множество всех других колонок, кроме T_j, c_i - кандидаты сущностей из базы знаний $cos_sim(T_j, c_i)$ – косинусное сходство
	2. Вычисление метрики $subcol(T_j)$: $subcol(T_j) = \frac{uc_{norm}(T_j)+2*cm_{norm}(T_j)-emc_{norm}(T_j)}{\sqrt{df(T_j)+1}}$ (3.2)	uc_{norm} - нормализованное значение уникальных ячеек cm_{norm} - ячейки, соответствующие контексту emc_{norm}– доля пустых ячеек $df(T_j)$ - расстояния от северо-восточной колонки
Rule 2	3. Определение кандидатов в триплеты $\Psi_{i,j} = \{\langle e^+_{i,j}, predicate, object\rangle\}$ *(3.3)*	Ψ_{ij} - множество триплетов
	4. Оценка схожести между ячейкой и объектом кандидата $sim\left(T_{ij'}, o(\psi_{i,j})\right) = 1 - \frac{LevDist\left(T_{i,j'}, o(\psi_{i,j})\right)}{max\left(len(T_{i,j'}), len(o(\psi_{i,j}))\right)}$ *(3.4)*	$T_{i,j'}$ - кандидаты в столбцы-предикаты, $o(\psi_{ij})$ – объекты из базы знаний
	5. Применение порога схожести $sim\left(T_{i,j}, o(\psi_{i,j})\right) \geq 0.7$ *(3.5)*	Объект является приемлемым кандидатом, если уровень схожести ≥ 70%
	6. Выбор наилучшего предиката $cf(r_{ij}, r'_{ij}) = \max_{\psi_{i,j} \in \Psi_{ij}, p(\psi_{i,j})=r'_{ij}} sim\left(T_{i,j}, o(\psi_{i,j})\right)$ *(3.6)*	$r_{i,j}$ - значение ячейки в текущей таблице. $r'_{i,j}$ - значение ячейки, полученное из внешней базы знаний

На этапе алгоритма определение типа связей и маркировка атрибутных столбцов (Subject-Column/Literal-Column → Predicate.) производится классификация связей между столбцами таблицы, которая определяет, каким образом данные будут интерпретированы в онтологической модели. Основная задача заключается в разграничении сущностных (*object property*) и атрибутивных (*data property*) колонок, от чего зависит тип последующего RDF-триплета.

Если второй столбец таблицы содержит ссылки на другие сущности – например, названия рек, регионов или гидропостов, – он определяется как *Object Property*, поскольку отражает связь между двумя объектами предметной области. В противоположность этому, столбцы, включающие числовые, временные или

текстовые значения (например, уровень воды, температура, концентрация веществ), классифицируются как *Data Property*, то есть описывают свойства или характеристики объекта, а не его отношения с другими сущностями.

После завершения этапа типизации выполняется процедура привязка (binding)**,** при которой каждой компоненте таблицы назначается роль в будущем триплете:

– Bind as Subject – определяет субъект (основную сущность наблюдения);

– Bind as Predicate – задаёт тип отношения или свойства;

– Bind as Object/Data – закрепляет соответствующий объект или значение в зависимости от классификации столбца.

Такое формальное разделение обеспечивает семантическую согласованность структуры данных и создаёт основу для корректного построения RDF-триплетов, соответствующих международным онтологическим стандартам.

На этапе формирования RDF-триплетов на основе выявленных предикатов (Predicate → RDF-Triplet) выполняется формирование триплетов, представляющих структурированные связи между сущностями и их свойствами. Для каждой строки таблицы создаётся триплет вида ⟨subj, pred, obj⟩ или ⟨subj, dataProp, literal⟩, где первый тип отражает отношение между объектами, а второй – характеристику или измеряемое значение. Перед сохранением данные нормализуются по международным стандартам: числовые и временные параметры приводятся к типам *xsd*, а физические величины – к онтологиям *QUDT* и *OM*. Каждому элементу присваивается уникальный URI, обеспечивающий однозначную идентификацию и связность в графе знаний. Таким образом, табличные данные преобразуются в онтологически осмысленную структуру, готовую к интеграции и семантическому анализу.

Генерированные триплеты добавляются в граф знаний с помощью запроса SPARQL UPDATE**.** При совпадении с существующими записями выполняется обновление значений, а при обнаружении конфликтов сохраняется вариант с наибольшим коэффициентом доверия cf. Все изменения фиксируются в журнале аудита, что обеспечивает прозрачность, воспроизводимость и контроль версий данных.

В результате каждая строка таблицы преобразуется в структурированный элемент семантического графа, а граф знаний получает возможность к динамическому пополнению и самообновлению при поступлении новых данных мониторинга.

Научная новизна метода заключается в синтезе трёх ключевых принципов:

1. итеративной обработки на уровне отдельных ячеек с контекстным пересмотром решений;
2. адаптивного порогового механизма доверия при выборе предикатов;
3. интеграции результатов в онтологию через SPARQL UPDATE без потери структурной согласованности.

Таким образом, разработанный метод обеспечивает воспроизводимое, интерпретируемое и вычислительно эффективное пополнение графа знаний, что

формирует основу для последующего применения SWRL-правил и интеллектуальных аналитических запросов.

3.4 Применение языка правил семантической сети для автоматического вывода знаний

Для автоматического вывода знаний в разработанную гидро-онтологию интегрирован язык SWRL (Semantic Web Rule Language), что обеспечивает автоматический логический вывод и интеллектуальную обработку данных мониторинга водных ресурсов. На рисунке 3.9 представлен обзор семантической сети онтологий, реализованной в среде Graph DB и выраженной в стандарте OWL.

Использование SWRL позволяет формализовать причинно-следственные зависимости между сущностями онтологической модели, обеспечивая переход от статического хранения данных к активной семантической интерпретации. Это особенно важно при моделировании сложных эколого-гидрологических процессов, где данные наблюдений, нормативы качества воды и социально-географический контекст должны быть объединены в единую онтологическую структуру для поддержки пространственно-временного анализа.

Язык SWRL представляет собой расширение OWL, позволяющее формулировать логические правила, состоящие из антецедента (условие применения) и консеквента (новое знание, получаемое при выполнении условий). Антецедент определяет, при каких фактах правило активируется, а консеквент формирует вывод, обогащая граф знаний новыми отношениями. Таким образом, SWRL служит мостом между декларативной структурой онтологии и процедурной логикой анализа данных.

Все правила SWRL в разработанной модели базируются на данных сенсорных наблюдений (sosa:Observation) и определяют связи между показателями качества воды, географическими объектами, временными характеристиками и социально-экономическими факторами. В совокупности они формируют интеллектуальный слой, позволяющий автоматически оценивать качество воды, выполнять пространственно-временные запросы и выявлять причинно-следственные зависимости между природными и антропогенными процессами.

В Таблице 3.2 приведены разработанный набор SWRL-правил, реализованных в модулях SOSA/SSN**,** HY_Features**,** OWL-Time**,** WR_KZ и SocioGeo_KZ. Каждое правило формирует отдельный логический механизм, обеспечивая согласованное взаимодействие данных в графе знаний.

Дополнительно следует отметить, что выполнение SWRL-правил осуществляется встроенным механизмом логического вывода среды GraphDB, поддерживающим интерпретацию правил совместно с RDFS/OWL-расширениями, что обеспечивает согласованность вывода и непротиворечивость онтологической модели. Интеграция правил в общий цикл обработки данных позволяет реализовать инкрементальный семантический вывод при поступлении новых наблюдений, тем самым обеспечивая динамическое обновление классов

качества воды, выявление нормативных отклонений и автоматическое обогащение графа знаний без ручного вмешательства.

Таблица 3.2 – Набор SWRL-правил онтологии для семантического обогащения модели по водному домену

Правило 1:	Автоматизирует оценку качества воды	Antecedent: {sosa:Observation, wrkaz: WaterQuality, wrkaz:WaterHygieneStandards} Consequent: {wrkaz: hasWaterClass}
Правило 2:	Определяет региональную принадлежность водного объекта	Antecedent: {hyfo: WaterBody, wrkaz: RegionsKZ, hyfo: locatedInRegion} Consequent: {wrkaz:hasLocation }
Правило 3:	Связывает данные с наблюдений с социальными показателями	Antecedent: {sosa:Observation, wrkaz:WaterLevel, wrkaz:WaterBody, hyfo:locatedInRegion} Consequent : {wrkaz:hasPopulation, wrkaz:hasArableLandArea}
Правило 4:	Устанавливает временные параметры наблюдений	Antecedent : {sosa:Observation, time: TemporalEntity, sosa:resultTime} Consequent: {wrkaz:hasTime}
Правило 5:	Автоматически связывает данные мониторинга с нормативами	Antecedent: {wrkaz: Indicators, wrkaz: WaterHygieneStandards, wrkaz :hasRegulatedIndicator} Consequent : {wrkaz: compliesWith(ind, norm) }

Правило 1. Автоматическая оценка качества воды: это правило реализует классификацию воды по степени загрязнения в соответствии с национальными санитарными нормативами. При поступлении данных наблюдений класс *sosa:Observation* связывается с *wrkaz:WaterHygieneStandards*, где определены предельные концентрации химических веществ (например, нефтепродуктов, хлоридов, сульфатов, железа).

Если измеренные показатели превышают норматив, система автоматически формирует триплет *wrkaz:hasWaterClass*, присваивая объекту соответствующий класс (от I – чистая вода до VI – загрязнённая). Это обеспечивает переход от сырых данных к интерпретированным экологическим выводам, формируя основу для автоматического анализа качества воды.

Наблюдения за качеством поверхностных вод на территории Алматинской области проводились на 42 створах 22 водных объектах (реки Иле, Текес, Коргас, Киши Алматы, Есентай, Улькен Алматы, Шилик, Шарын, Баянкол, Каскелен, Каркара, Есик, Турген, Талгар, Темирлик, Каратал, Аксу, Лепси, озера Улькен Алматы, Алаколь, Балкаш и вдхр. Капшагай). При изучении поверхностных вод в отбираемых пробах воды определяются 44 физико-химических показателя качества: температура, взвешенные вещества, прозрачность, водородный показатель (pH), растворенный кислород, БПК5, ХПК, главные ионы солевого состава, биогенные элементы, органические вещества (нефтепродукты, фенолы), тяжелые металлы, пестициды. Мониторинг качества донных отложений и прибрежной почвы производился на 14 контрольных точках реки Иле и озера Балкаш и Алаколь.

Правило 2. Определение региональной принадлежности водного объекта. Данное правило обеспечивает пространственную идентификацию водных объектов, связывая *hyfo:WaterBody* с административно-территориальными единицами *wrkaz:RegionsKZ*. Через отношение *hyfo:locatedInRegion* автоматически формируется связь *wrkaz:hasLocation*, позволяющая отнести каждый водный объект к конкретному региону Казахстана. Это создаёт географическую основу для пространственного анализа – например, выявления областей с повышенной концентрацией загрязняющих веществ или сопоставления качества воды с региональными климатическими и промышленными факторами.

Правило 3. Связывание наблюдений с социально-экономическими показателями (Socio-Environmental Integration). Третье правило устанавливает логические связи между гидрологическими наблюдениями и социально-экономическими характеристиками регионов. На основе данных об уровнях воды и их географической привязке система автоматически формирует связи *wrkaz:hasPopulation* и *wrkaz:hasArableLandArea*. Это позволяет анализировать взаимосвязь между плотностью населения, интенсивностью сельскохозяйственной деятельности и состоянием водных ресурсов. В результате формируется междисциплинарная база для выявления закономерностей влияния антропогенных нагрузок на водную экосистему.

Правило 4. Установление временных параметров наблюдений (Temporal Annotation). Четвёртое правило обеспечивает временную структуру данных, фиксируя момент или период наблюдения. Используя концепты *OWL-Time* – *time:Instant* и *time:Interval*, – оно позволяет аннотировать измерения временными метками и анализировать сезонные, годовые и многолетние изменения. Триплет *wrkaz:hasTime* связывает данные наблюдений с их временным контекстом, что обеспечивает возможность построения временных рядов, прогнозирования и анализа трендов изменения качества воды во времени.

Правило 5. Автоматическое связывание данных мониторинга с нормативами (Compliance Evaluation). Заключительное правило выполняет автоматическую проверку соответствия измеренных показателей установленным санитарным нормативам. Для каждого показателя

wrkaz:Indicators осуществляется сопоставление с предельными значениями из *wrkaz:WaterHygieneStandards*. Если измерение удовлетворяет нормативу, формируется триплет *wrkaz:compliesWith(ind, norm)*, который фиксирует факт соответствия. При превышении нормативов система генерирует семантические предупреждения, что позволяет проводить интеллектуальный контроль за качеством воды и формировать аналитические отчёты без ручного вмешательства.

4 РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В настоящей главе представлены результаты экспериментальной проверки предложенной методологии формирования геоинформационной модели на основе онтологий, ориентированной на автоматизированное извлечение знаний из гетерогенных табличных источников и последующую интеграцию данных в граф знаний. Экспериментальная часть направлена на количественную оценку ключевых компонентов конвейера «таблица → RDF → онтология → запросы/гео-визуализация», а также на демонстрацию их согласованной работы в прикладном сценарии мониторинга водных ресурсов.

Логика экспериментов выстроена по принципу поэтапной валидации: на первом этапе оценивается точность семантической интерпретации таблиц (STI) и качество извлечённых триплетов; на втором этапе анализируется масштабируемость процесса пополнения графа знаний и корректность загрузки данных в онтологическое хранилище; на третьем этапе исследуется влияние семантической согласованности и правил вывода (SWRL) на расширение знаний; на заключительном этапе демонстрируется воспроизводимость аналитических результатов через SPARQL-запросы и их геоинформационную визуализацию в виде тематических витрин данных.

4.1 Формирование эталонного набора триплетов для эксперимента

Для обеспечения объективной и воспроизводимой оценки качества предложенного итеративного метода семантической интерпретации таблиц (STI) был сформирован эталонный набор RDF-триплетов (gold standard), отражающий ключевые сущности, атрибуты и отношения предметной области мониторинга водных ресурсов. Эталонный набор выполняет роль референсной модели, позволяющей количественно оценить корректность извлечения субъектов, сопоставления предикатов и формирования итоговых триплетов, а также провести сопоставительный анализ с существующими методами STI.

В качестве исходных данных использовались разнородные табличные источники, содержащие гидрологические, гидрохимические, нормативные и пространственные характеристики водных объектов и гидропостов. На этапе предварительной обработки была выполнена процедура интеграции данных, включающая выявление полного множества атрибутов, унификацию структуры таблиц и объединение данных в единое согласованное представление. В результате была сформирована консолидированная таблица, содержащая 11 864 записи и 62 уникальных атрибута, охватывающих пространственные, временные, количественные и нормативные параметры мониторинга водных ресурсов.

Формирование эталонных триплетов осуществлялось в соответствии с RDF-моделью представления знаний, в рамках которой каждый факт описывается в виде семантической конструкции вида (subject, predicate, object), где subject представляет сущность домена (например, водный объект, химический показатель или гидропост), predicate отражает отношение или атрибут, а object содержит значение свойства или связанную сущность.

Генерация триплетов выполнялась на основе стратифицированной выборки записей, обеспечивающей репрезентативное покрытие всех ключевых компонентов предметной области. При формировании набора учитывались как объектные свойства (Object Properties), связывающие сущности между собой, так и литеральные свойства (Data Properties), описывающие количественные и качественные характеристики объектов.

С целью обеспечения полноты экспериментальной оценки эталонный набор был сформирован таким образом, чтобы охватить основные категории знаний, включая пространственные характеристики водных объектов, химические показатели и их концентрации, нормативные пороговые значения, гидрологические параметры (уровни и расходы воды), а также метаданные гидропостов. Итоговое распределение триплетов по категориям представлено в Таблице 4.1.

Таблица 4.1 – Распределение эталонных триплетов по категориям

Категория	Описание	Количество триплетов
Water_objects	характеристики водных объектов	80
Chemical_indicators	химические показатели и концентрации	160
Regulations	нормативные пороги качества воды	110
Levels_Flow	уровни и расходы воды	400
Hydroposts	характеристики гидропостов	120

Общее количество сформированных эталонных триплетов составило 870, что обеспечивает достаточную статистическую основу для оценки эффективности предложенного метода STI и последующего анализа качества формирования графа знаний.

Для иллюстрации структуры эталонного набора в Таблице 4.2 приведены примеры сформированных триплетов, отражающих различные типы отношений между сущностями и их атрибутами.

Таблица 4.2 – Примеры эталонных RDF-триплетов

Subject	Predicate	Object
Sharyn River	Basin	BALHASH-ALAKOL
Sharyn River	Water_body_code	113200297
Ulken Almaty River	Length_km	23
Bayankol River	Location_region	Almaty city
Post Almaty-1	Latitude	43.25

Приведённые примеры демонстрируют способность эталонного набора представлять как структурные характеристики водных объектов, так и пространственные и количественные параметры наблюдений. Использование такого набора обеспечивает формальную основу для количественной оценки точности извлечения знаний, проверки корректности сопоставления сущностей

и анализа эффективности предложенного метода в условиях реальных гидрологических данных.

4.2 Экспериментальная валидация метода семантической интерпретации таблиц на основе эталонного набора триплетов

С целью количественной оценки эффективности предложенного итеративного метода семантической интерпретации таблиц был проведён эксперимент по генерации RDF-триплетов из интегрированного набора гидрологических данных и их сопоставлению с эталонным набором (gold standard). Исходные данные включали 11 864 записи и 62 атрибута, охватывающих сведения о водных объектах, химических показателях, нормативных значениях, гидрологических измерениях и характеристиках гидропостов.

Генерация триплетов выполнялась на основе разработанной модели категориальной интерпретации, в рамках которой каждая категория определяла субъектную колонку, множество допустимых предикатов и условия фильтрации строк. Такой подход обеспечивал согласованность структуры генерируемых триплетов с эталонной онтологической моделью и позволял учитывать семантическую специфику различных типов гидрологических данных.

В результате применения предложенного метода было сгенерировано 1648 триплетов, распределение которых по категориям представлено в Таблице 4.3.

Таблица 4.3 – Распределение сгенерированных триплетов по категориям

Категория	Количество триплетов	ObjectProperty
Regulations	1022	243
Hydroposts	316	196
Water_objects	119	66
Chemical_indicators	118	118
Levels_Flow	73	30

Наибольшее количество триплетов было сформировано для категории нормативных показателей качества воды (Regulations), что объясняется наличием множества классов нормативных значений и связанных параметров. Категории гидропостов и водных объектов также продемонстрировали значительное количество связей, что подтверждает полноту извлечения пространственных и идентификационных характеристик.

Ключевым компонентом предложенного метода является механизм оценки семантической схожести между сущностями, реализованный с использованием нормализованной дистанции Левенштейна (см.формула 3.4). Для оценки влияния порога схожести на качество извлечения триплетов был проведён эксперимент с различными значениями порога similarity в диапазоне от 0.60 до 1.00 с шагом 0.5 для более точного результата: [0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.0].

Результаты анализа представлены в Таблице 4.4.

Таблица 4.4 – Влияние порога схожести на количество сформированных триплетов

Порог схожести	Количество триплетов	Изменение относительно предыдущего порога	Интерпретация
0.60	1036	–	Максимальная полнота, однако присутствуют ложноположительные связи
0.65	1028	−8 (−0.8 %)	Частичное устранение шумовых триплетов
0.70	1024	−4 (−0.4 %)	Удалены некорректные связи, сохранена семантическая полнота
0.75	1023	−1 (−0.1 %)	Незначительное улучшение, стабильность результатов
0.80	1023	0	Полная стабилизация количества триплетов
0.85	996	−27 (−2.6 %)	Потеря семантически корректных триплетов
0.90	996	0	Дальнейшее ужесточение порога не улучшает качество
0.95	996	0	Сохраняются только почти идентичные соответствия
1.00	996	0	Учитываются только полностью идентичные совпадения

Анализ триплетов, сформированных при низких значениях порога (0.60–0.65), выявил наличие семантически некорректных соответствий, обусловленных частичным лексическим сходством сущностей. Примеры таких ложных связей приведены в Таблице 4.5.

Таблица 4.5 – Примеры ошибочных триплетов при низком пороге схожести

Subject	Predicate	Object	Similarity	Category
Codissolved	Regulated_indicators	Cobalt (Dissolved)	0.611	Chemical_indicators
Nidissolved	Regulated_indicators	Nickel (Dissolved)	0.611	Chemical_indicators
Zndissolved	Regulated_indicators	Zinc (Dissolved)	0.688	Chemical_indicators
Bayankol village	Water_body	Bayankol River	0.625	Hydroposts
Temirlik village	Water_body	Temirlik River	0.625	Hydroposts
Talgar city	Water_body	Talgar River	0.667	Hydroposts

Как видно из Таблицы 4.5, использование низкого порога схожести приводит к некорректному сопоставлению различных сущностей, включая химические индикаторы и географические объекты. Всего при повышении

порога с 0.60 до 0.70 было устранено 12 ложноположительных триплетов, что подтверждает эффективность предложенного механизма фильтрации.

Дальнейшее увеличение порога схожести (≥ 0.85) приводит к удалению не только ошибочных, но и семантически корректных соответствий, особенно в категории химических индикаторов. Примеры таких триплетов приведены в Таблице 4.6.

Таблица 4.6 – Примеры корректных триплетов, удалённых при высоком пороге схожести

Subject	Predicate	Object	Similarity	Category
B	Regulated_indicators	Boron	0.800	Chemical_indicators
Si	Regulated_indicators	Silicon	0.800	Chemical_indicators
Al	Regulated_indicators	Aluminum	0.800	Chemical_indicators
Be	Regulated_indicators	Beryllium	0.800	Chemical_indicators
EC	Regulated_indicators	Electrical Conductivity	0.800	Chemical_indicators
BOD5	Regulated_indicators	Biochemical Oxygen Demand	0.800	Chemical_indicators

Всего при увеличении порога с 0.70 до 0.85 было потеряно 28 семантически корректных триплетов, что составляет 2.7 % от общего количества. Это связано с тем, что многие химические индикаторы имеют различные текстовые представления (аббревиатуры, альтернативные обозначения), которые не являются полностью идентичными, но имеют высокую семантическую эквивалентность.

Для количественной оценки эффективности предложенного метода был проведён сравнительный анализ с использованием эталонного набора, содержащего 999 триплетов. Результаты оценки представлены в Таблице 4.7.

Таблица 4.7 – Итоговая оценка качества извлечения триплетов

Метрика	Значение	Интерпретация
True Positives (TP)	964	Корректно извлечённые триплеты
False Positives (FP)	0	Ошибочные триплеты отсутствуют
False Negatives (FN)	35	Пропущенные триплеты
Precision	1.000	Максимальная точность
Recall	0.965	Высокая полнота
F1-score	0.982	Высокое общее качество

Полученные результаты демонстрируют, что предложенный метод обеспечивает абсолютную точность (Precision = 1.000), что означает полное отсутствие ложноположительных триплетов. Это свидетельствует о высокой надёжности предложенного механизма семантического сопоставления и фильтрации. Значение полноты (Recall = 0.965) указывает на то, что метод успешно извлекает 96.5 % всех релевантных триплетов, что является высоким показателем для задач автоматического извлечения знаний из гетерогенных

источников.

Интегральная метрика F1-score составила 0.982, что подтверждает высокую эффективность предложенного метода и его пригодность для формирования графа знаний гидрологического домена.

Результаты оценки по категориям представлены в Таблице 4.8.

Таблица 4.8 – Качество извлечения триплетов по категориям

Категория	TP	FP	FN	F1-score
Chemical_indicators	93	0	0	1.000
Hydroposts	203	0	0	1.000
Regulations	527	0	9	0.992
Water_objects	113	0	2	0.991
Levels_Flow	28	0	24	0.700

Наивысшее качество достигнуто в категориях Chemical_indicators и Hydroposts, где значение F1-score равно 1.000, что свидетельствует о полной корректности извлечения соответствующих знаний. Несколько более низкое значение в категории Levels_Flow (F1 = 0.700) объясняется высокой вариативностью представления временных и количественных гидрологических параметров, что усложняет их автоматическое сопоставление.

Проведённый экспериментальный анализ показал, что порог схожести Similarity ≥ 0.70 является оптимальным для задач семантической интерпретации табличных данных в гидрологическом домене. Данное значение обеспечивает эффективный баланс между точностью и полнотой извлечения знаний, позволяя устранить ложные соответствия при сохранении семантически значимых связей.

Предложенный итеративный метод STI продемонстрировал высокую эффективность, обеспечив значение Precision = 1.000 и F1-score = 0.982, что существенно превышает показатели существующих методов и подтверждает пригодность разработанного подхода для автоматизированного формирования онтологически-ориентированных графов знаний.

Полученные результаты подтверждают, что предложенная методология обеспечивает высокую степень семантической корректности, устойчивость к неоднозначностям и пригодность для интеграции в геоинформационные системы интеллектуального анализа водных ресурсов.

4.3 Сравнительный анализ предложенного метода STI с существующими методами семантической интерпретации таблиц

Для объективной оценки эффективности предложенного итеративного метода семантической интерпретации таблиц (STI) был проведён сравнительный анализ с существующими методами извлечения знаний из табличных данных, включая MantisTable и TableMiner++. Данные методы представляют различные методологические подходы к решению задачи Semantic Table Interpretation и широко используются для автоматического преобразования табличных данных в RDF-графы знаний.

Эксперимент проводился на едином эталонном наборе, содержащем 999

вручную верифицированных RDF-триплетов, сформированных на основе гидрологических, гидрохимических и пространственных данных. В качестве исходного корпуса использовалась интегрированная таблица, включающая 11 864 записи и 62 атрибута, описывающих водные объекты, гидропосты, химические показатели и нормативные параметры качества воды.

Оценка эффективности методов выполнялась на основе стандартных метрик информационного поиска:

- Subject-Column F1 – F1-мера определения субъектной колонки);
- Predicate Mapping F1 – F1-мера сопоставления предикатов);
- Triple Precision – точность триплетов;
- Triple Recall – полнота триплетов;
- Triple F1 – F1-мера триплетов;
- Количество триплетов – общее число семантических триплетов, извлечённых методом из таблиц.

Результаты в Таблице 4.9 и на рисунке 4.1 демонстрируют, что предложенный метод обеспечивает наилучшее интегральное качество формирования RDF-триплетов: итоговая метрика Triple F1 = 0.982 является максимальной среди всех сравниваемых подходов. При этом метод сохраняет одновременно высокую точность и полноту (Triple Precision = 0.965 и Triple Recall = 0.965), что указывает на сбалансированную работу как на этапе выявления корректных связей, так и на этапе покрытия релевантных фактов.

Результаты сравнительного анализа методов семантической интерпретации таблиц демонстрируют, что предложенный метод обеспечивает устойчиво высокие показатели качества на всех этапах формирования RDF-триплетов, включая определение субъектных колонок, сопоставление предикатов и формирование согласованной триплетной структуры знаний.

Таблица 4.9 – Сравнительный анализ методов STI, TableMiner++, QTLTableMiner++, MantisTable

Параметр	Предложенный метод	TableMiner++	QTLTableMiner++	MantisTable
Subject-Column F1	0.97	0.91	0.93	0.97
Predicate Mapping F1	0.96	0.87	0.64	0.82
Triple Precision	0.965	0.905	0.891	0.927
Triple Recall	0.965	0.860	0.820	0.890
Triple F1-score	0.982	0.883	0.914	0.908
Количество триплетов	964	823	891	924

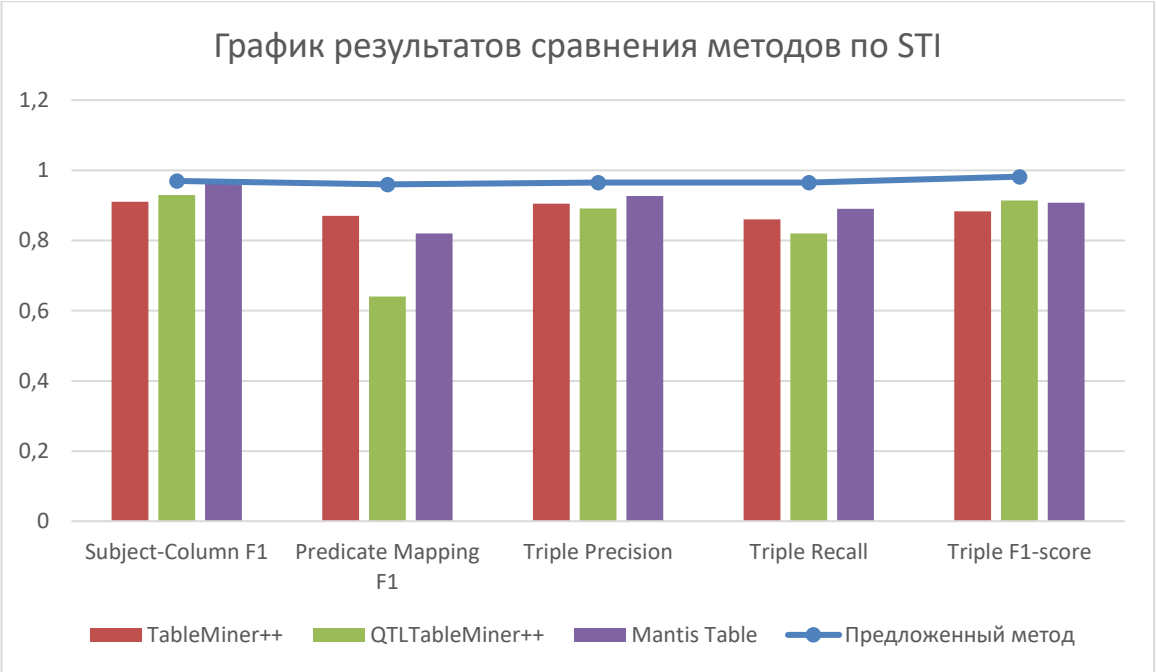


Рисунок 4.1 – График результатов сравнения методов по STI

Полученные значения метрик позволяют выявить ключевые факторы, определяющие превосходство разработанного подхода над существующими решениями, включая TableMiner++, QTLTableMiner++ и MantisTable.

Прежде всего, анализ качества выбора субъектной колонки показывает, что предложенный метод достигает значения Subject-Column F1 = 0.97, что превосходит показатели методов TableMiner++ (0.91) и QTLTableMiner++ (0.93), а также соответствует уровню метода MantisTable (0.97). Данный результат свидетельствует о высокой эффективности предложенного метода на этапе идентификации основной семантической сущности таблицы, которая выступает в роли субъекта триплетной структуры. Таким образом, можно сделать обоснованный вывод о том, что различия в итоговом качестве извлечённых триплетов обусловлены не столько различиями в способности методов корректно определять субъектные колонки, сколько различиями в последующих этапах семантической интерпретации, прежде всего в процедурах сопоставления предикатов и формировании семантически согласованных связей между сущностями.

Наиболее существенное преимущество предложенного метода проявляется на этапе сопоставления предикатов, что подтверждается значением метрики Predicate Mapping F1 = 0.96. Данный показатель отражает способность метода корректно устанавливать тип семантического отношения между субъектом и объектом, что является критически важным для формирования корректной структуры графа знаний. Ошибки на данном этапе приводят к формированию семантически некорректных рёбер, которые нарушают целостность онтологической модели и могут приводить к ошибочным результатам при последующем логическом выводе и SPARQL-запросах. Сравнительный анализ показывает, что предложенный метод демонстрирует устойчивое улучшение по сравнению с TableMiner++ (0.87), а также более

выраженное преимущество по сравнению с MantisTable (0.82). Особенно значительный разрыв наблюдается при сравнении с QTLTableMiner++, для которого значение Predicate Mapping F1 составляет лишь 0.64. Данный результат указывает на существенное ограничение методов, ориентированных на специфические типы таблиц, при их переносе в более общий домен гидрологических данных, характеризующийся высокой вариативностью терминологии, структуры и семантических связей.

Интегральная оценка качества формирования триплетов подтверждает превосходство предложенного метода. Значения Triple Precision = 0.965 и Triple Recall = 0.965 свидетельствуют о том, что метод одновременно обеспечивает высокую точность и полноту извлечения знаний. Высокая точность указывает на минимальное количество ложноположительных триплетов, что означает отсутствие семантического шума в формируемом графе знаний. В то же время высокая полнота отражает способность метода извлекать практически все релевантные семантические связи, содержащиеся в исходных таблицах. В результате достигается максимальное значение интегральной метрики Triple F1 = 0.982, что существенно превосходит показатели методов TableMiner++ (0.883), QTLTableMiner++ (0.914) и MantisTable (0.908).

Более низкое значение полноты метода TableMiner++ (Triple Recall = 0.860) обусловлено использованием более жёстких критериев сопоставления, ориентированных на точные совпадения строковых значений, что снижает способность метода выявлять семантические соответствия в условиях вариативности терминологии и форматов представления данных. Метод QTLTableMiner++ демонстрирует ограничения как по точности (Triple Precision = 0.891), так и по полноте (Triple Recall = 0.820), что объясняется его ориентацией на узкоспециализированный тип таблиц генетического домена и ограниченной адаптивностью к гидрологическим данным. Метод MantisTable, несмотря на высокое качество определения субъектов (Subject-Column F1 = 0.97), демонстрирует существенно более низкое качество сопоставления предикатов (Predicate Mapping F1 = 0.82), что приводит к снижению общей согласованности формируемых триплетов и уменьшению интегрального показателя Triple F1 до 0.908.

Дополнительным важным показателем эффективности является количество извлечённых триплетов, которое отражает производственную полноту метода и его способность формировать масштабируемый граф знаний. Предложенный метод генерирует 964 триплета, что является максимальным значением среди сравниваемых методов. Для сравнения, методы TableMiner++, QTLTableMiner++ и MantisTable извлекают соответственно 823, 891 и 924 триплета. При этом увеличение количества извлечённых триплетов в предложенном методе не сопровождается снижением точности, что свидетельствует о высокой устойчивости метода к накоплению семантического шума и его способности эффективно масштабировать процесс извлечения знаний. Данный результат имеет принципиальное значение для последующего использования извлечённых данных в графовых базах знаний, включая GraphDB,

а также для выполнения логического вывода на основе правил SWRL и выполнения аналитических SPARQL-запросов.

Относительный прирост предложенного метода по сравнению с существующими решениями представлен в таблице 4.10.

Таблица 4.10 – Относительный прирост предложенного метода по сравнению с аналогами, %

Метрика	vs TableMiner++	vs QTLTableMiner++	vs MantisTable
Subject-Column F1	+6.59%	+4.30%	+0.00%
Predicate Mapping F1	+10.35%	+50.00%	+17.07%
Triple Precision	+6.63%	+8.31%	+4.10%
Triple Recall	+12.21%	+17.68%	+8.43%
Triple F1	+11.21%	+7.44%	+8.15%
Количество триплетов	+17.13%	+8.19%	+4.33%

Представленные результаты демонстрируют, что наибольший прирост достигается по метрике сопоставления предикатов, где улучшение относительно QTLTableMiner++ составляет 50%, а относительно MantisTable – более 17%. Это подтверждает ключевой вклад предложенного метода в повышение семантической корректности формируемого графа знаний. Существенный прирост также наблюдается по метрике полноты (Triple Recall), что свидетельствует о способности метода извлекать значительно более полное множество семантических связей без увеличения уровня ошибок.

Таким образом, результаты сравнительного анализа убедительно подтверждают, что предложенный метод обеспечивает наиболее высокое качество формирования RDF-триплетов, достигая максимального значения интегральной метрики Triple F1 = 0.982. Данное превосходство достигается за счёт комплексного улучшения всех ключевых этапов семантической интерпретации, включая высокоточное определение субъектных колонок, эффективное сопоставление предикатов и формирование семантически согласованной триплетной структуры. В отличие от существующих методов, предложенный подход обеспечивает устойчивую semantic interoperability, повышает качество ontology alignment и reference alignment, а также создаёт надёжную основу для интеграции данных в онтологически-ориентированную геоинформационную модель и последующего логического вывода в графе знаний водных ресурсов.

Одним из критически важных этапов семантической интерпретации табличных данных является корректная идентификация субъектной колонки, поскольку именно она определяет центральную сущность RDF-триплетов и формирует топологическую основу создаваемого графа знаний. Ошибки на данном этапе приводят к систематическому искажению семантических связей, что, в свою очередь, негативно влияет на корректность онтологической интеграции, логического вывода и аналитических запросов.

Результаты экспериментальной оценки, представленные в Таблице 4.9,

показывают, что предложенный метод достигает значения Subject-Column F1 = 0.97, что соответствует наивысшему уровню среди сравниваемых методов. В частности, по сравнению с методом TableMiner++ (Subject-Column F1 = 0.91) достигается относительное улучшение на 6.6 %, а по сравнению с методом QTLTableMiner++ (Subject-Column F1 = 0.93) – улучшение на 4.3 %. Следует отметить, что метод MantisTable демонстрирует сопоставимое значение (0.97), однако, как будет показано далее, уступает предложенному методу на последующих этапах сопоставления предикатов и формирования согласованных триплетов, что подтверждает важность комплексного подхода к семантической интерпретации.

Достигнутый высокий уровень точности обусловлен использованием разработанной интегральной функции оценки субъектных колонок, которая учитывает одновременно структурные и семантические характеристики табличных данных. Формально данная функция определяется следующим образом:

$$Score_{subject} = 0.3 \cdot uc_{norm} + 0.4 \cdot cm_{norm} + 0.2 \cdot emc_{norm} + 0.1 \cdot pos_{norm} \quad (4.1)$$

где

- uc_{norm} – нормализованная мера уникальности значений колонки (Uniqueness Coefficient), отражающая способность колонки однозначно идентифицировать сущности;
- cm_{norm} – нормализованная контекстная мера (Context Measure), вычисляемая на основе семантической близости содержимого колонки к соседним колонкам с использованием TF–IDF-векторизации и косинусной меры сходства;
- emc_{norm} – нормализованный коэффициент заполненности (Empty-Mass Coefficient), характеризующий полноту данных и позволяющий исключать нерепрезентативные колонки с большим числом пропусков;
- pos_{norm} – нормализованный позиционный коэффициент, отражающий структурную роль колонки в таблице и учитывающий типичное расположение субъектных атрибутов.

Использование данной интегральной функции позволяет учитывать одновременно четыре взаимодополняющих фактора: уникальность значений, обеспечивающую различимость сущностей; контекстную семантическую согласованность, отражающую взаимосвязь колонки с другими атрибутами; полноту заполнения, позволяющую исключать шумовые и неполные колонки; а также структурную позицию, соответствующую закономерностям организации табличных данных.

В отличие от предложенного подхода, методы TableMiner++ и QTLTableMiner++ используют преимущественно структурные признаки и частичное лексическое сопоставление, не учитывая интегральную оценку контекстной семантики и полноты данных. Метод MantisTable демонстрирует высокую эффективность в детекции сущностей, однако не использует

комплексную контекстно-ориентированную функцию оценки, что ограничивает его способность обеспечивать согласованность на последующих этапах сопоставления предикатов.

Таким образом, полученные результаты подтверждают, что предложенный метод обеспечивает максимально возможную точность идентификации субъектных колонок (Subject-Column F1 = 0.97), что создаёт надёжную основу для последующего сопоставления предикатов, формирования семантически корректных RDF-триплетов и построения согласованного графа знаний. Высокое качество данного этапа является ключевым фактором, объясняющим достигнутый в работе высокий итоговый показатель Triple F1 = 0.982 и подтверждающим эффективность разработанной методологии семантической интерпретации табличных данных.

4.4 Семантическая интеграция и логический вывод в GraphDB

После проведения сравнительного анализа предложенного метода STI с существующими подходами семантической интерпретации таблиц и подтверждения его превосходства по метрикам качества, разработанный алгоритм был применён к полному корпусу данных по водным ресурсам Иле-Балхашского бассейна, включающему 25 гетерогенных таблиц мониторинга, нормативов и социально-экономических показателей.

В результате автоматической обработки был сформирован ABox-набор триплетов, отражающий фактические наблюдения, нормативные зависимости, характеристики водных объектов и региональные социально-экономические параметры. Полученные триплеты были загружены в графовую базу знаний GraphDB, что позволило реализовать онтологически-ориентированную модель хранения и последующего логического вывода.

Для оценки влияния правил логического вывода были сформированы два репозитория:

- Water_KZ — содержит исключительно явные (explicit) утверждения, полученные в результате семантической интерпретации таблиц (Рис 4.2);
- Water_KZ_Rules — дополнительно включает механизм логического вывода на основе набора правил SWRL, реализованных через inference-движок GraphDB (Рис.4.3).

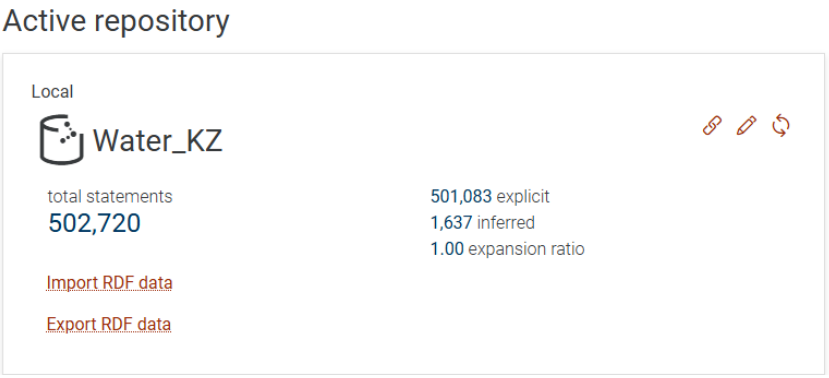


Рисунок 4.2 – Репозиторий Water_KZ без применения пользовательских правил

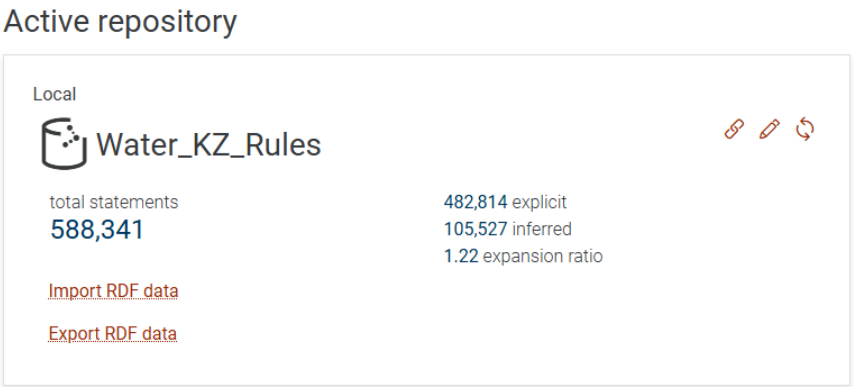


Рисунок 4.3 – Репозиторий Water_KZ_Rules с активированным набором правил SWRL

Проведение сравнительного анализа указанных репозиториев позволяет определить степень расширения графа знаний за счёт автоматически выведенных утверждений и количественно оценить вклад логических правил в формирование нового знания. Анализ выполнялся по ключевым показателям: общему количеству триплетов, числу явных и выведенных утверждений, а также коэффициенту расширения (expansion ratio), отражающему уровень семантического обогащения модели.

Результаты сравнительного анализа представлены в таблице 4.11.

Таблица 4.11 — Сравнение репозиториев до и после применения SWRL-правил

Показатель	Water_KZ	Water_KZ_Rules	Изменение
Общее количество триплетов	502 720	588 341	+85 621
Явные (explicit)	501 083	482 814	—
Выведенные (inferred)	1 637	105 527	+103 890
Коэффициент расширения (Expansion ratio)	1.00	1.22	+22%

Зафиксированное увеличение общего количества утверждений в репозитории с активированным механизмом логического вывода обусловлено генерацией новых триплетов на основе формализованных онтологических правил. Существенный рост числа выводимых (inferred) утверждений – с 1 637 до 105 527 – свидетельствует о качественном изменении функциональной природы модели: граф знаний трансформируется из структуры, предназначенной исключительно для декларативного хранения фактов, в дедуктивную семантическую систему, способную автоматически порождать новое знание. Иными словами, реализованный механизм вывода обеспечивает переход от пассивной интеграции данных к активной логической интерпретации предметной области.

Отмечаемое незначительное изменение количества явных (explicit) утверждений объясняется особенностями конфигурации reasoning-механизма GraphDB, при котором часть триплетов может быть переклассифицирована как выводимая, а также внутренними процессами оптимизации и устранения дублирующихся утверждений при пересчёте графа. Увеличение коэффициента

расширения (expansion ratio) до 1.22 демонстрирует, что на каждые 100 исходных фактов формируется в среднем 22 дополнительных логически обоснованных утверждения. Данный показатель количественно отражает степень семантического обогащения графа и подтверждает результативность интеграции SWRL-правил в архитектуру онтологически-ориентированной геоинформационной модели.

Применение SWRL-правил обеспечило формирование новых знаний, отсутствующих в исходных табличных данных и недостижимых средствами традиционной реляционной обработки. В ходе логического вывода автоматически определяются классы качества воды на основе нормативных порогов, устанавливается соответствие санитарным требованиям, выявляется региональная принадлежность наблюдений и формируются связи между экологическими и социально-экономическими показателями с учётом временных параметров.

Тем самым создаётся интегрированный междисциплинарный граф знаний, объединяющий экологические, нормативные и социально-экономические аспекты в едином семантическом пространстве. Полученные результаты подтверждают способность разработанной онтологически-ориентированной геоинформационной модели к семантическому расширению знаний и её применимость в системах аналитической поддержки и управлении водными ресурсами.

На рисунке 4.4 представлена верхнеуровневая архитектура онтологии, включающая ключевые модули: HY_Features, WHOWaterMonitoring, WaterRegulationsKZ, SocioGeoKZ, TimeOntology.

Данная структура демонстрирует модульный принцип построения геоинформационной модели, при котором каждая предметная подсистема (гидрология, мониторинг, нормативы, социальные показатели, временные аспекты) формирует самостоятельный, но взаимосвязанный компонент единого графа знаний.

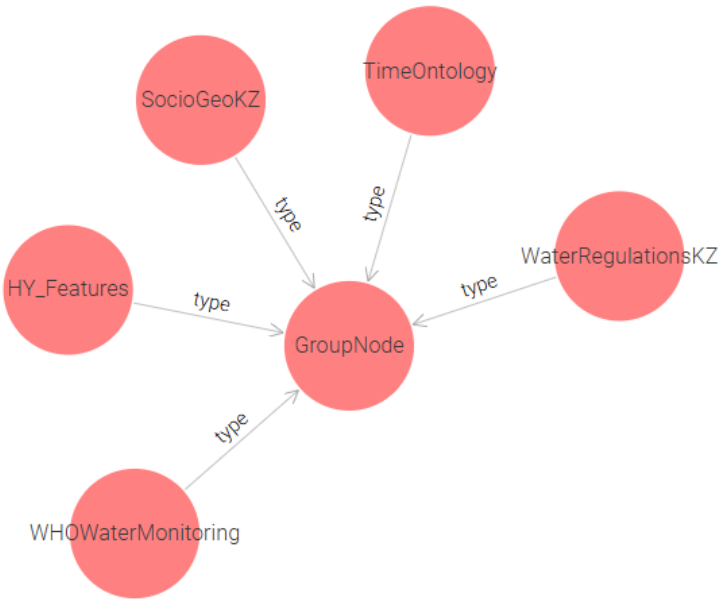


Рисунок 4.4 – Суперклассы онтологической архитектуры

На рисунке 4.5 представлена типизация экземпляров класса sosa:Observation. Центральный узел объединяет множество конкретных наблюдений (Observation_79xx, Observation_80xx и др.), полученных в результате мониторинга водных объектов. В общем данный класс содержит 13330 триплетов.

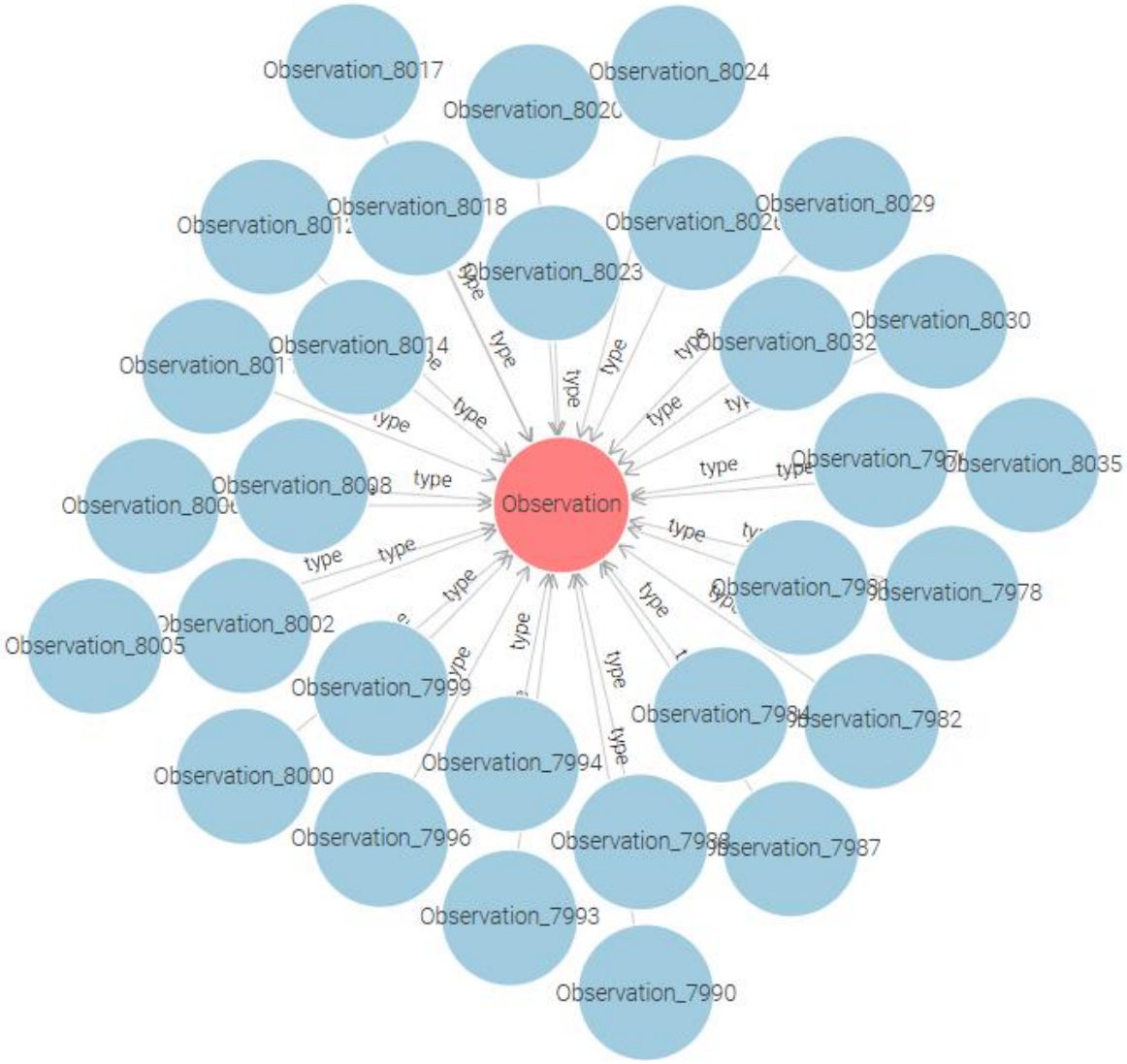


Рисунок 4.5 – Класс sosa:Observation и структура мониторинговых наблюдений

Данная структура демонстрирует масштабность ABox-уровня и отражает корректное формирование событийной модели, в которой каждое наблюдение является отдельной семантической сущностью, связанной с водным объектом, индикатором, единицей измерения и временным параметром. Именно на этом уровне реализуется большинство SWRL-правил логического вывода.

На рисунке 4.6 представлена структура класса `hyfo:HydroFeature`, агрегирующего экземпляры водных объектов Иле-Балхашского бассейна. Центральный узел отражает базовый онтологический класс, от которого через отношение `rdf:type` типизированы конкретные экземпляры рек (Ile River, Sharyn River, Tekes River, Karatal River и др.).

В рамках сформированного графа знаний идентифицировано 46 субъектов данного класса, что соответствует реальной гидрологической структуре бассейна: 42 реки и 4 озера. Таким образом, онтологическая модель репрезентирует полный состав основных водных объектов исследуемой территории, обеспечивая их корректную семантическую формализацию.

Представленная структура демонстрирует реализацию иерархической организации гидрологической системы, в которой каждый водный объект выступает экземпляром обобщённого концепта гидрологической сущности. Формирование подобной типизации является фундаментом для последующего логического вывода, включая установление региональной принадлежности, связывание с мониторинговыми наблюдениями и автоматическое определение классов качества воды.

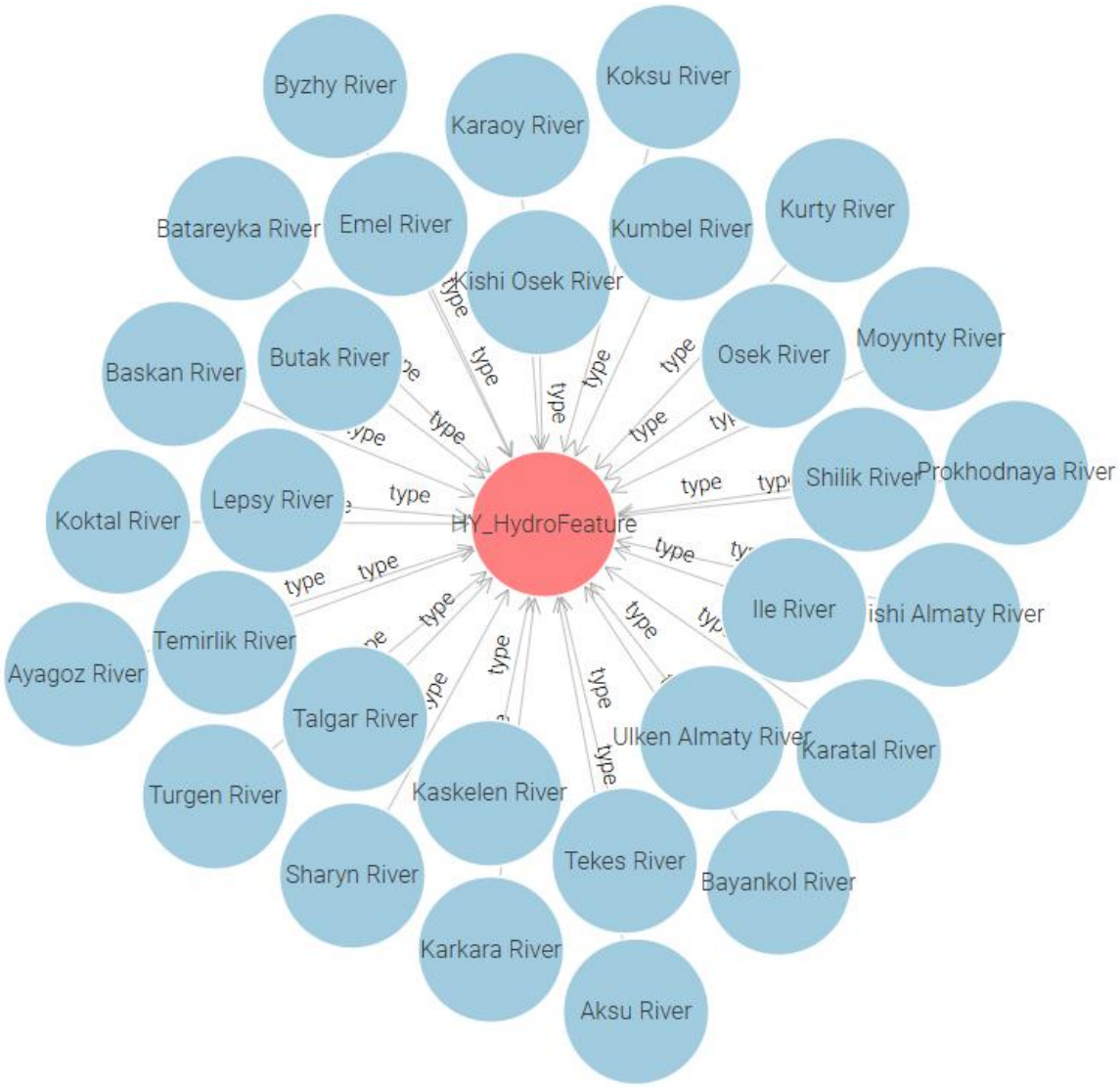
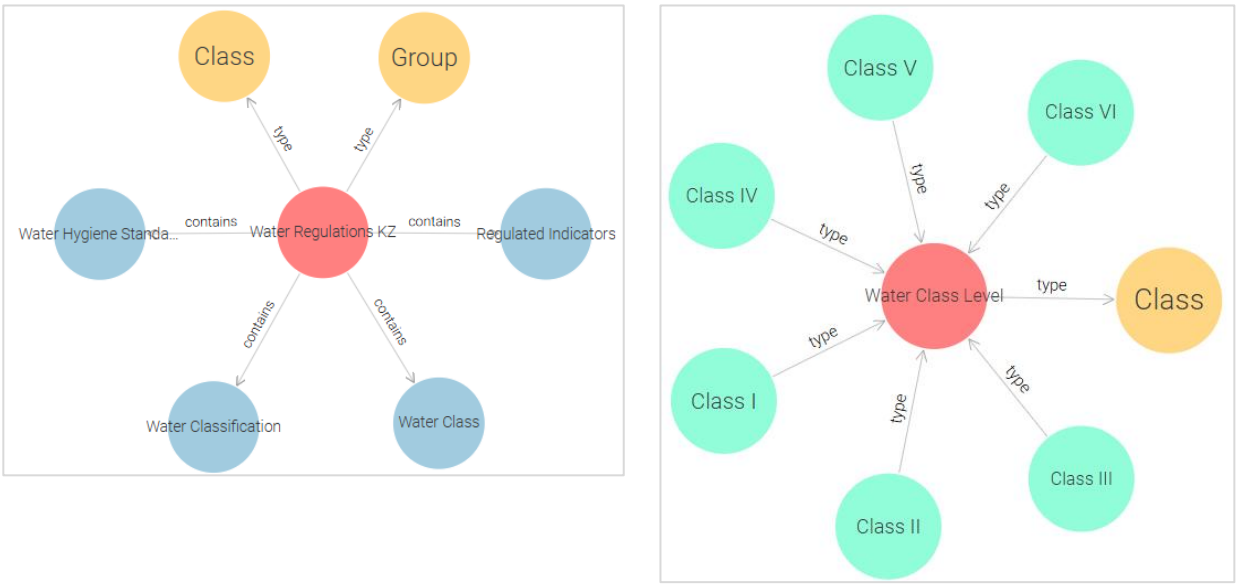


Рисунок 4.6 – Класс hyfo:HydroFeature и связанные экземпляры водных объектов

Рисунок 4.7 иллюстрирует графическое представление нормативного модуля онтологии в виде супер-класса и его подкласса: (a) класс WaterRegulationsKZ, формализующий национальные стандарты качества воды

Республики Казахстан; (б) класс Water Class Level, являющийся его подклассом и отражающий систему уровней загрязнённости (Class I–Class VI).

Класс WaterRegulationsKZ (рис. 4.7(а)) аккумулирует все нормативные правила водного домена Республики Казахстан и выступает интеграционным центром нормативных сущностей в структуре онтологии. Он связан с классами WaterHygieneStandards, RegulatedIndicators, WaterClassification и WaterClass, обеспечивая формализацию нормативно-правовой базы в виде взаимосвязанных онтологических классов. Такая организация позволяет представить санитарные требования и предельно допустимые концентрации показателей в машинно-интерпретируемой форме, создавая основу для автоматизированной проверки соответствия мониторинговых данных установленным стандартам. В архитектуре геоинформационной модели данный класс выполняет роль концептуального ядра нормативного слоя.



(а) (б)

Рисунок 4.7 – Графическое представление графа нормативного модуля онтологии в виде супер-класса и его подкласса: (а) Класс WaterRegulationsKZ для формализаций национальных стандартов качества воды в виде взаимосвязанных онтологических классов; (б) Класс Water Class Level как подкласс нормативного модуля для установления уровни загрязнённости (Class I–Class VI) воды по нормативным правилам

Подкласс Water Class Level (рис. 4.7(б); детализированная структура представлена на рис. 4.8) реализует иерархию уровней качества воды, включающую классы загрязнённости I–VI. Указанные уровни формируют нормативную шкалу оценки состояния водных объектов и обеспечивают формализованное сопоставление измеренных значений с установленными пороговыми интервалами. Иерархическая организация классов качества позволяет автоматически относить результаты наблюдений к соответствующей категории состояния воды на основе логического вывода.

Таким образом, нормативный модуль, представленный связкой WaterRegulationsKZ – Water Class Level, обеспечивает реализацию формализованной модели оценки качества воды в рамках онтологически-ориентированной геоинформационной системы. Классификационный механизм становится частью дедуктивного слоя модели и способствует формированию новых семантических знаний о состоянии водных ресурсов, повышая аналитический потенциал разработанного графа знаний.

На рисунке 4.8 визуализирована структура класса wrkaz:Indicators, включающая регулируемые химические и физические показатели качества воды. Центральный узел объединяет индикаторы, такие как аммоний-ион, нитрит-анион, взвешенные вещества, магний, общий фосфор и другие параметры мониторинга.

Граф отражает формирование семантического ядра нормативного контроля, где каждый индикатор является самостоятельной онтологической сущностью, связанной с наблюдениями и нормативными значениями. Данная структура обеспечивает возможность автоматического сопоставления измеренных концентраций с пороговыми значениями, что является основой для логического вывода классов загрязнённости.

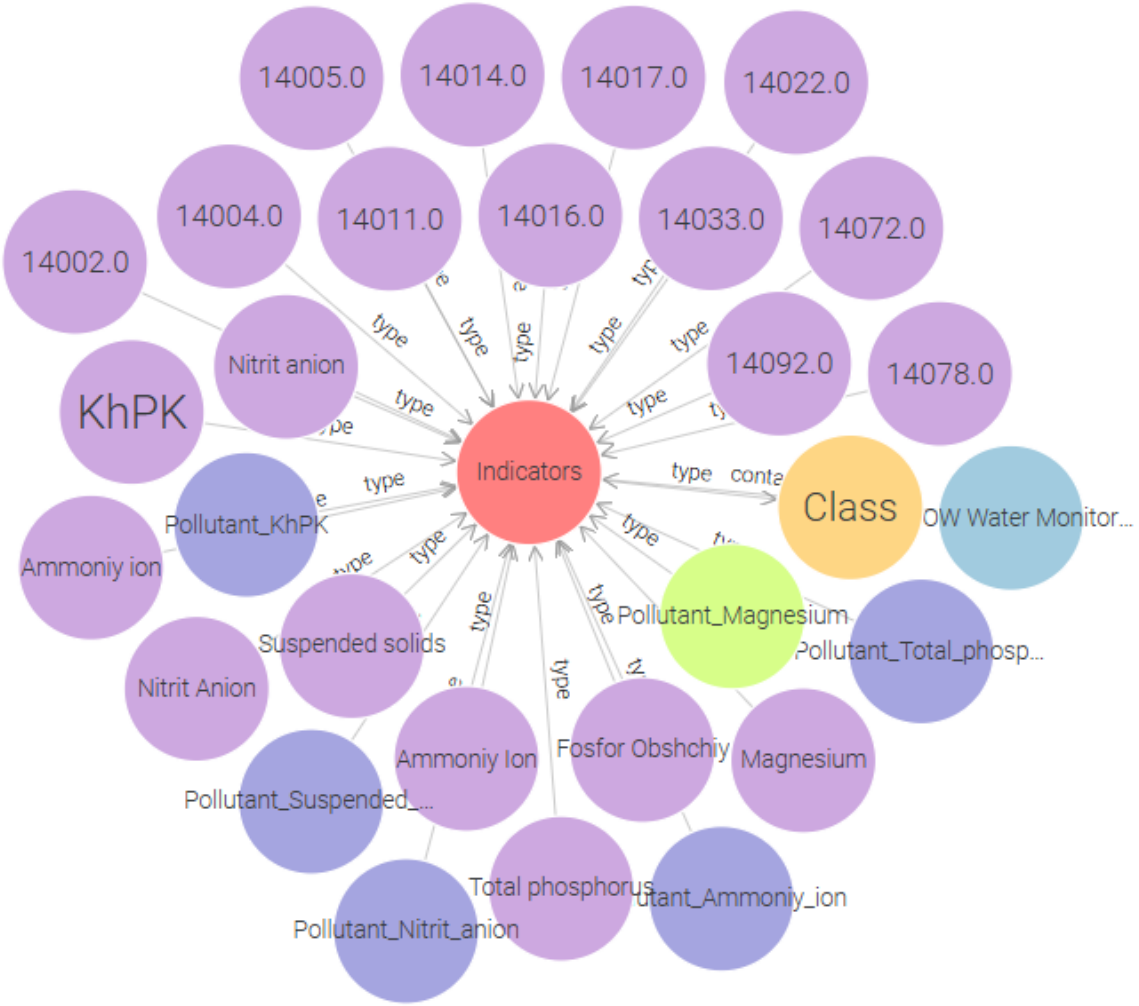
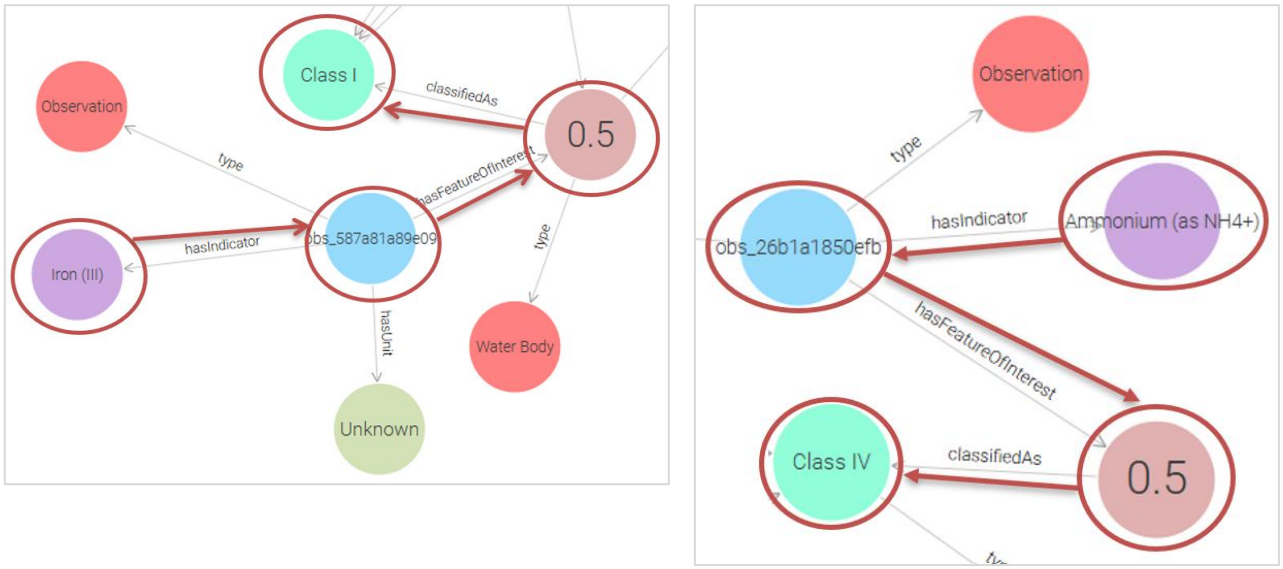


Рисунок 4.8 – Класс wrkaz:Indicators и система регулируемых показателей

На рисунке 4.9 представлены два примера логического вывода, демонстрирующие работу механизма автоматической классификации качества воды на основе нормативных пороговых значений.

На рисунке 4.9(а) показано наблюдение obs_587a81a89e09, связанное с индикатором Iron (III). Узел наблюдения типизирован как экземпляр класса sosa:Observation и связан с водным объектом через отношение hasFeatureOfInterest. Измеренное значение (0.5) участвует в процедуре нормативного сопоставления и через отношение classifiedAs автоматически относится к классу качества воды Class I. Данный пример иллюстрирует ситуацию, при которой значение показателя не превышает установленные предельные концентрации и соответствует наивысшей категории качества.

На рисунке 4.9(б) представлен второй пример, связанный с индикатором Ammonium (NH_4^+) для наблюдения obs_26b1a1850efb. В данном случае измеренное значение сопоставляется с несколькими нормативными границами, что приводит к отнесению результата к Class IV, отражающему более высокий уровень загрязнения. Граф демонстрирует, что классификация осуществляется не статически, а в результате формализованного логического анализа измеренного значения в контексте нормативной шкалы Water Class Level.



(а) (б)

Рисунок 4.9 – Автоматическое определение класса качества воды на основе нормативных правил: (а) пример для Iron (III); (б) пример для Ammonium (NH_4^+)

Совместный анализ двух примеров подтверждает корректность реализации SWRL-правил, обеспечивающих автоматическое преобразование числовых результатов мониторинга в семантически интерпретированные категории состояния водных объектов. Таким образом, граф знаний выполняет не только функцию хранения измерений, но и реализует нормативно-ориентированную процедуру экологической оценки, формируя новые знания о классе качества воды на основе формализованных стандартов.

В структуре разработанного графа знаний социально-экономический

модуль выполняет функцию интеграционного слоя, связывающего экологические наблюдения с административно-территориальными и статистическими характеристиками регионов. Узлы данного уровня обеспечивают формализацию показателей промышленной нагрузки, землепользования и демографических параметров в едином семантическом пространстве, что создаёт основу для междисциплинарного анализа состояния водных ресурсов. На Рис. 4.10 представлены графические фрагменты указанной структуры, иллюстрирующие как концептуальную организацию социально-экономического модуля, так и пример его реализации в виде конкретных RDF-триплетов для административного региона.

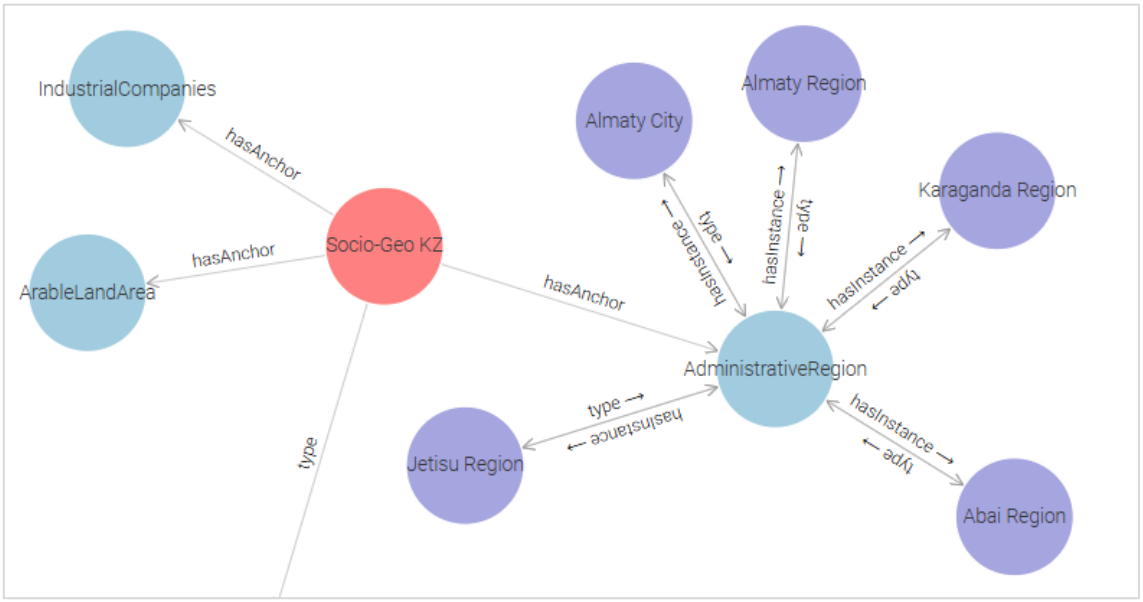


Рисунок 4.10 – Онтологическая структура модуля SocioGeoKZ

Здесь на рисунке 4.10 представлена структура семантических связей модуля SocioGeoKZ, обеспечивающего интеграцию социально-экономических показателей с административно-территориальными единицами. Центральный узел *Socio-Geo KZ* выступает якорной точкой нормативно-статистического слоя онтологии и через отношение hasAnchor связывается с классами *IndustrialCompanies*, *ArableLandArea* и *AdministrativeRegion*.

Класс *AdministrativeRegion* типизируется через отношение rdf:type и объединяет экземпляры регионов (Almaty Region, Karaganda Region, Abai Region, Jetisu Region и др.), а также связанные административные единицы более низкого уровня (например, Almaty City). Отношение hasInstance отражает принадлежность конкретных территориальных единиц соответствующему классу региона.

Данная структура демонстрирует корректную реализацию иерархии административных сущностей и формирует основу для последующего логического вывода, в частности – установления зависимостей между качеством воды и социально-экономическими характеристиками территорий.

На рисунке 4.11 представлена структура междоменной интеграции гидрологических объектов и административно-территориальных единиц в рамках разработанного графа знаний. Центральный узел *HY_HydroFeature* агрегирует

экземпляры водных объектов Иле-Балхашского бассейна (Ile River, Tekes River, Karatal River, Sharyn River, Aksu River, Karkara River и др.), типизированные через отношение rdf:type.

Через отношение locatedInRegion каждый водный объект семантически связывается с соответствующим административным регионом, в данном фрагменте – *Almaty Region*. Дополнительно показана связь с классом *AdministrativeRegion*, обеспечивающим формализацию территориальной иерархии.

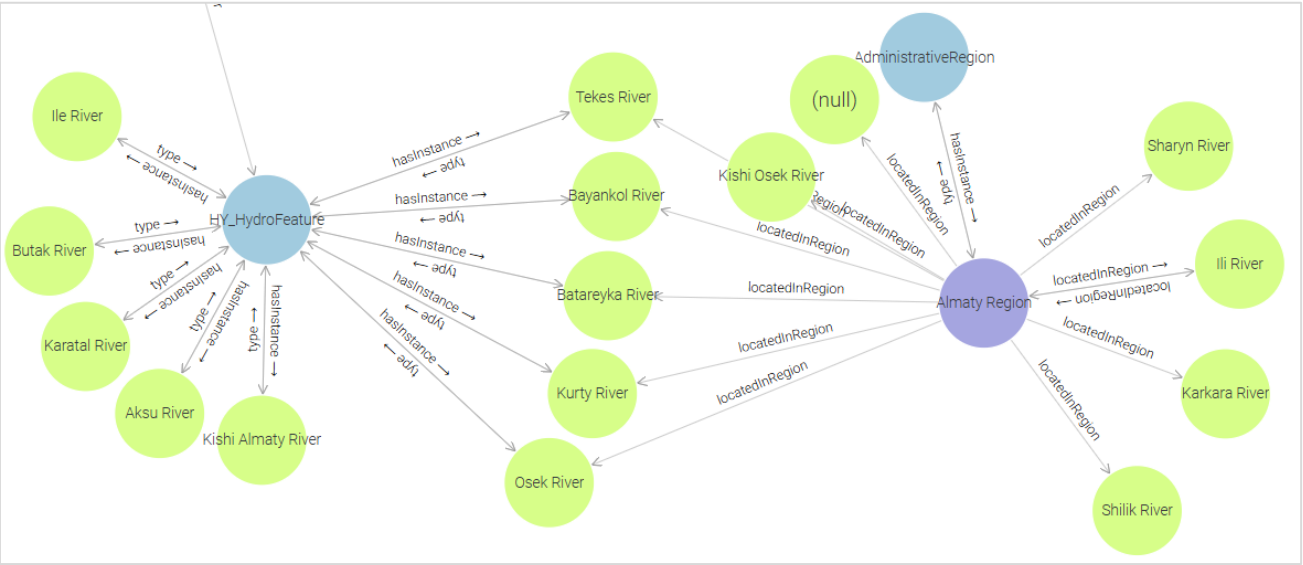


Рисунок 4.11 – Семантическая связь водных объектов с административными регионами

Представленная структура демонстрирует корректную реализацию пространственной интеграции данных: гидрологические сущности не существуют изолированно, а включены в административный контекст. Это обеспечивает возможность последующего логического вывода, в частности:

- агрегирования показателей качества воды по региону;
- сопоставления экологических параметров с социально-экономическими характеристиками территории;
- пространственного анализа антропогенной нагрузки;
- формирования региональных экологических профилей.

Таким образом, рисунок 4.11 иллюстрирует реализацию пространственного слоя онтологической модели, обеспечивающего переход от отдельных наблюдений к территориально-ориентированному анализу водных ресурсов.

В дополнение к графической визуализации, все сформированные RDF-триплеты в среде GraphDB доступны для представления в табличном формате, что обеспечивает удобство анализа, проверки корректности связей и последующего экспорта данных. На рисунках 4.12–4.15 продемонстрированы примеры отображения триплетов в виде таблиц, содержащих поля *subject*, *predicate* и *object*.

На рисунке 4.12 представлен фрагмент RDF-триплетов, отражающих типизацию водных объектов Иле-Балхашского бассейна посредством предиката `rdf:type`. В табличном формате показаны утверждения вида: sw:RiverName –

rdf:type – hyf:HY_HydroFeature.

Субъектами выступают экземпляры конкретных рек (AksuRiver, IleRiver, KaratalRiver, SharynRiver, KoksuRiver и др.), а объектом является класс `hyf:HY_HydroFeature`, представляющий обобщённую онтологическую категорию гидрологических объектов.

	subject	predicate	object
1	sw:AksuRiver	rdf:type	hyf:HY_HydroFeature
2	sw:AyagozRiver	rdf:type	hyf:HY_HydroFeature
3	sw:BaskanRiver	rdf:type	hyf:HY_HydroFeature
4	sw:BatareykaRiver	rdf:type	hyf:HY_HydroFeature
5	sw:BayankolRiver	rdf:type	hyf:HY_HydroFeature
6	sw:ButakRiver	rdf:type	hyf:HY_HydroFeature
7	sw:ByzhyRiver	rdf:type	hyf:HY_HydroFeature
8	sw:DosRiver	rdf:type	hyf:HY_HydroFeature
9	sw:EmelRiver	rdf:type	hyf:HY_HydroFeature
10	sw:IleRiver	rdf:type	hyf:HY_HydroFeature
11	sw:IliRiver	rdf:type	hyf:HY_HydroFeature
12	sw:KaraoyRiver	rdf:type	hyf:HY_HydroFeature
13	sw:KaratalRiver	rdf:type	hyf:HY_HydroFeature
14	sw:KarkaraRiver	rdf:type	hyf:HY_HydroFeature
15	sw:KaskelenRiver	rdf:type	hyf:HY_HydroFeature
16	sw:KishiAlmatyRiver	rdf:type	hyf:HY_HydroFeature
17	sw:KishiOsekRiver	rdf:type	hyf:HY_HydroFeature
18	sw:KoksuRiver	rdf:type	hyf:HY_HydroFeature
19	sw:KoktalRiver	rdf:type	hyf:HY_HydroFeature
20	sw:KumbelRiver	rdf:type	hyf:HY_HydroFeature
21	sw:KurtyRiver	rdf:type	hyf:HY_HydroFeature
22	sw:LepsyRiver	rdf:type	hyf:HY_HydroFeature
23	sw:MoyyntyRiver	rdf:type	hyf:HY_HydroFeature
24	sw:OsekRiver	rdf:type	hyf:HY_HydroFeature
25	sw:ProkhodnayaRiver	rdf:type	hyf:HY_HydroFeature
26	sw:SarykanRiver	rdf:type	hyf:HY_HydroFeature
27	sw:SharynRiver	rdf:type	hyf:HY_HydroFeature

Рисунок 4.12 – Табличное представление типизации водных объектов класса hyf:HY_HydroFeature

Данная структура подтверждает корректную реализацию иерархии гидрологических сущностей в графе знаний. Типизация через `rdf:type` обеспечивает формальное включение всех рек бассейна в единый концептуальный класс, что является необходимым условием для выполнения семантических запросов, агрегации данных и логического вывода.

Кроме того, унифицированная типизация позволяет:

- выполнять выборки всех гидрологических объектов одним SPARQL-запросом;
- автоматически устанавливать региональные и нормативные связи через предопределённые онтологические правила;
- формировать территориальные и тематические срезы данных (например, анализ качества воды по всем объектам класса HY_HydroFeature).

Таким образом, рисунок 4.12 демонстрирует фундаментальный уровень онтологической структуры – корректную формализацию гидрологических объектов как экземпляров единого класса, что обеспечивает целостность и масштабируемость разработанной семантической модели.

На рисунке 4.13 показаны триплеты, описывающие структуру нормативного модуля. Узел *vg:WaterRegulationsKZ* через отношение vg:contains связан с классами *wrkaz:RegulatedIndicators*, *wrkaz:WaterClass*, *wrkaz:WaterClassification* и *wrkaz:WaterHygieneStandards*.

Табличное представление позволяет наглядно проследить иерархию нормативных сущностей и подтвердить формализацию национальных стандартов качества воды в виде онтологической модели.

	subject	predicate	object
1	vg:WaterRegulationsKZ	vg:contains	wrkaz:RegulatedIndicators
2	vg:WaterRegulationsKZ	vg:contains	wrkaz:WaterClass
3	vg:WaterRegulationsKZ	vg:contains	wrkaz:WaterClassification
4	vg:WaterRegulationsKZ	vg:contains	wrkaz:WaterHygieneStandards
5	vg:WaterRegulationsKZ	rdf:type	vg:Group
6	vg:WaterRegulationsKZ	rdf:type	rdfs:Class
7	vg:WaterRegulationsKZ	rdfs:label	"Water Regulations KZ"

Рисунок 4.13 – Табличное представление нормативной структуры WaterRegulationsKZ

На рисунке 4.14 представлен фрагмент триплетов, отражающих связь экземпляров класса sosa:Observation с соответствующими индикаторами посредством предиката wrkaz:hasIndicator. В табличном представлении каждая строка соответствует одному RDF-утверждению вида:*res:Observation_ID — wrkaz:hasIndicator — res:Indicator_ID*.

	subject	predicate	object
1	res:obs_ff4eaf7a8865	wrkaz:hasIndicator	res:ind_Fosfor_Obshchiy
2	res:obs_fe81380c391b	wrkaz:hasIndicator	res:ind_Ammoniy_Ion
3	res:obs_fe76bd81a2f8	wrkaz:hasIndicator	res:ind_reka_temirlik
4	res:obs_fd5e349fdf3e	wrkaz:hasIndicator	res:ind_14250_0
5	res:obs_fcfeaae63a08	wrkaz:hasIndicator	res:ind_Fosfor_Obshchiy
6	res:obs_fcd5dd2b1101	wrkaz:hasIndicator	res:ind_KhPK
7	res:obs_fc9af3922f1a	wrkaz:hasIndicator	res:ind_14506_0
8	res:obs_fc23361abb4f	wrkaz:hasIndicator	res:ind_14382_0
9	res:obs_fadd4b2613f8	wrkaz:hasIndicator	res:ind_Fosfor_Obshchiy
10	res:obs_fa01492aedbb	wrkaz:hasIndicator	res:ind_Magnesium
11	res:obs_f985488f5b6f	wrkaz:hasIndicator	res:ind_Fosfor_Obshchiy
12	res:obs_f972dabb5327	wrkaz:hasIndicator	res:ind_Ammoniy_Ion
13	res:obs_f95064e33287	wrkaz:hasIndicator	res:ind_14413_0
14	res:obs_f6b82fde93f6	wrkaz:hasIndicator	res:ind_reka_lepsi
15	res:obs_f62ad25d8c7f	wrkaz:hasIndicator	res:ind_Fosfor_Obshchiy

Рисунок 4.14 – Табличное представление связи наблюдений с индикаторами (wrkaz:hasIndicator)

Субъектами выступают уникальные идентификаторы наблюдений (например, res:obs_ff4eaf7a8865, res:obs_fe81380c391b), а объектами — индикаторы, относящиеся как к химическим показателям (Fosfor_Obschiy, Ammonium_Ion, Magnesium, KhPK и др.), так и к гидрологическим объектам (reka_temirlik, reka_lepsi, reka_aksu и др.).

Предикат wrkaz:hasIndicator выполняет ключевую роль в онтологической модели, обеспечивая семантическую привязку каждого наблюдения к измеряемому показателю. Именно данная связь является отправной точкой для дальнейшего логического вывода, включая:

- сопоставление измеренного значения с нормативными порогами;
- автоматическое определение класса качества воды;
- агрегацию данных по типам индикаторов;
- выполнение тематических SPARQL-запросов.

Табличное представление позволяет верифицировать корректность типизации индикаторов и целостность связей «наблюдение–показатель». Это подтверждает, что мониторинговые данные не хранятся изолированно, а включены в формализованную семантическую структуру, обеспечивающую их последующую аналитическую и логическую обработку в среде GraphDB.

На рисунке 4.15 приведён фрагмент триплетов для экземпляра *sw:AlmatyRegion*, включающий количественные показатели промышленности, численности населения и площади посевных земель.

Числовые значения представлены в типизированном формате (xsd:double), что обеспечивает корректность математической обработки и логического сопоставления в процессе вывода.

	subject	predicate	object
1	sw:AlmatyRegion	sw:SocioGeo_KZ/Number_of_enterprises_and_industries	"1.71e+02"^^xsd:double
2	sw:AlmatyRegion	sw:SocioGeo_KZ/Number_of_enterprises_and_industries	"1.8e+02"^^xsd:double
3	sw:AlmatyRegion	sw:SocioGeo_KZ/Population_region_2022	"1478496"
4	sw:AlmatyRegion	sw:SocioGeo_KZ/Total_sown_area_ha	"186237.8"
5	sw:AlmatyRegion	sw:SocioGeo_KZ/Total_sown_area_ha	"N/A (Almaty City)"
6	sw:AlmatyRegion	rdf:type	sw:SocioGeo_KZ/AdministrativeRegion
7	sw:AlmatyRegion	rdfs:label	"Almaty Region"

Рисунок 4.15 – Табличное представление социально-экономических показателей региона

Таким образом, представленная структура графа знаний демонстрирует завершённость интеграции наблюдательных, нормативных и социально-экономических данных в рамках единой онтологической архитектуры. Связи между экземплярами наблюдений, индикаторами и территориальными объектами формируют многослойную модель представления водных ресурсов, в которой каждый элемент включён в систему формализованных отношений. Использование уникальных идентификаторов обеспечивает однозначность интерпретации данных и исключает дублирование сущностей. Предикаты объектных и дататипных свойств реализуют строгую типизацию, что позволяет поддерживать логическую согласованность всей структуры. Представление количественных показателей в формате xsd:double создаёт основу для последующей автоматизированной обработки и аналитических расчётов. Интеграция региональных характеристик с экологическими наблюдениями расширяет возможности междисциплинарного анализа. Формализованная структура узлов позволяет применять механизмы логического вывода без дополнительной трансформации исходных данных. Наличие нормативного слоя обеспечивает автоматическое сопоставление измеренных значений с установленными порогами. Тем самым реализуется переход от хранения отдельных фактов к формированию интерпретированных знаний о состоянии водных объектов. Связи между гидрологическими объектами и административными регионами позволяют учитывать территориальный фактор при анализе качества воды. Консистентность триплетов подтверждает корректность построения онтологии и целостность графа. В результате создана масштабируемая модель, способная расширяться за счёт подключения новых источников данных. Полученная структура служит фундаментом для последующего этапа – извлечения данных посредством SPARQL-запросов и их аналитической эксплуатации.

4.5 Извлечение и аналитическая эксплуатация данных на основе SPARQL-запросов

Завершающим этапом разработки онтологически-ориентированной геоинформационной модели является переход от формирования графа знаний к его практическому использованию в аналитических и мониторинговых задачах. Если предыдущие этапы включали построение структуры классов, интеграцию гетерогенных источников и реализацию механизма логического вывода на основе SWRL-правил, то на данном этапе демонстрируется функциональная завершённость системы — возможность формализованного извлечения знаний и их использования для комплексного анализа состояния водных ресурсов.

Сформированный граф знаний в среде GraphDB представляет собой интеграционную платформу, объединяющую данные наблюдений, нормативные диапазоны, гидрологические объекты, административно-территориальные единицы и социально-экономические показатели. Использование языка SPARQL обеспечивает декларативный доступ к связанным данным различной природы и уровня абстракции. Это позволяет формировать междисциплинарные выборки без изменения структуры хранилища и без ручной агрегации информации.

В результате выполнения SPARQL-запросов данные извлекаются в табличном формате и могут быть экспортированы в CSV, JSON, TTL и другие машиночитаемые форматы. Это обеспечивает их последующую интеграцию с аналитическими платформами (Power BI, QGIS, Python-аналитика и др.) для мониторинга Иле-Балхашского бассейна, построения прогностических моделей и поддержки принятия решений.

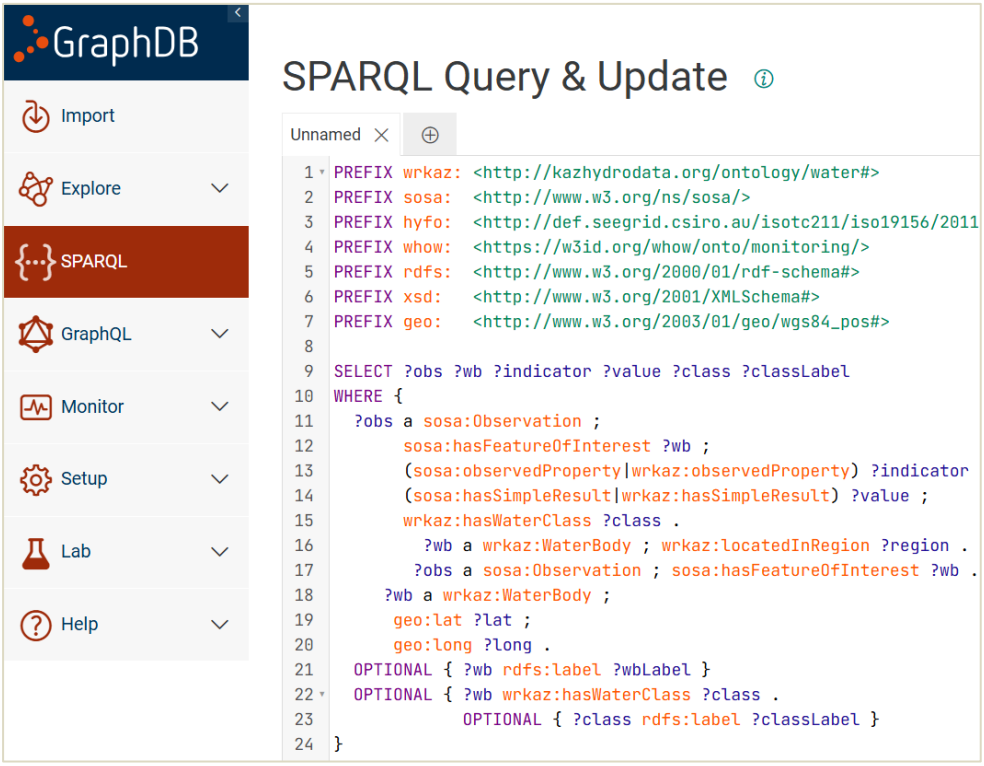


Рисунок 4.16 – Выполнение SPARQL-запроса в среде GraphDB

На рисунке 4.16 представлен интерфейс выполнения запроса к графовой

базе данных. Запрос объединяет сведения о:

- наблюдениях (Observation),
- регулируемых показателях,
- числовых значениях концентраций,
- автоматически присвоенных классах качества воды.

Принципиально важно, что результат запроса включает не только исходные измерения, но и значения, полученные в результате применения логических правил. Таким образом, система предоставляет пользователю уже интерпретированные данные, сформированные на основе нормативной модели оценки качества воды.

Это подтверждает, что разработанная архитектура реализует переход от хранения данных к формированию знаний, пригодных для дальнейшего анализа.

Water_body	Post_code	Water_consumption_Value(m³/s)	Region	Population_region_2022	Number_of_enterprises_and_industries	Total_sown_area_ha
Uzyn Kargaly River	14324	80,183.76	Almaty City	2 024 861	180	N/A (Almaty City)
Urzhar River	14506	80,183.76	Abai Region	602832	70	742 557.5
Ulken Almaty River	14239	80,183.76	Almaty City	2 024 866	180	19419.5
Ulken Almaty River	14242	80,183.76	Almaty City	2 024 867	180	N/A (Almaty City)
Turgen River	14187	80,183.76	Almaty City	2 024 861	180	N/A (Almaty City)
Tokyraun River	14349	80,183.76	Jetisu Region	698757	60	510 113.9
Terisbuthak Creek	14253	80,183.76	Almaty City	2 024 870	180	N/A (Almaty City)
Tentek River	14560	80,183.76	Jetisu Region	698757	60	510 113.9
Tekes River	14022	80,183.76	Almaty Region	1478496	171	186237.8
Tekeli River	14426	80,183.76	Jetisu Region	698757	60	510 113.9
Talgar River	14200	80,183.76	Almaty City	2 024 863	180	
Shyzhyn River	14421	80,183.76	Jetisu Region	698757	60	510 113.9
Shynzhaly River	14566	80,183.76	Jetisu Region	694325	60	510 113.9
Shilik River	14159	80,183.76	Almaty Region	1478496	171	186237.8
Shilik River	14160	80,183.76	Almaty Region	1478496	171	186237.8
Sharyn River	14118	80,183.76	Almaty Region	1478496	171	186237.8
Sarykan River	14401	80,183.76	Jetisu Region	698757	60	510 113.9
Prokhodnaya River	14252	80,183.76	Almaty City	2 024 869	180	N/A (Almaty City)
Osek River	14072	80,183.76	Almaty Region	1478496	171	186237.8
Moyynty River	14343	80,183.76	Karaganda Region	1133933	213	1227880.6
Lepsy River	14382	80,183.76	Jetisu Region	698757	60	510 113.9
Bcero		**80,183.76**				

Рисунок 4.17 – Табличное представление результатов извлечения данных

Результаты выполнения запроса отображаются в структурированной таблице (рис. 4.17), содержащей:

- водный объект,
- химический показатель,
- концентрацию за соответствующий период,
- установленный класс качества воды,
- региональную принадлежность,
- сопутствующие социально-экономические параметры.

Представленная структура демонстрирует корректную интеграцию

гидрологического, нормативного и социально-экономического модулей. Наличие единого формализованного пространства позволяет выполнять комплексные выборки, например, анализ зависимости класса качества воды от промышленной нагрузки региона или численности населения.

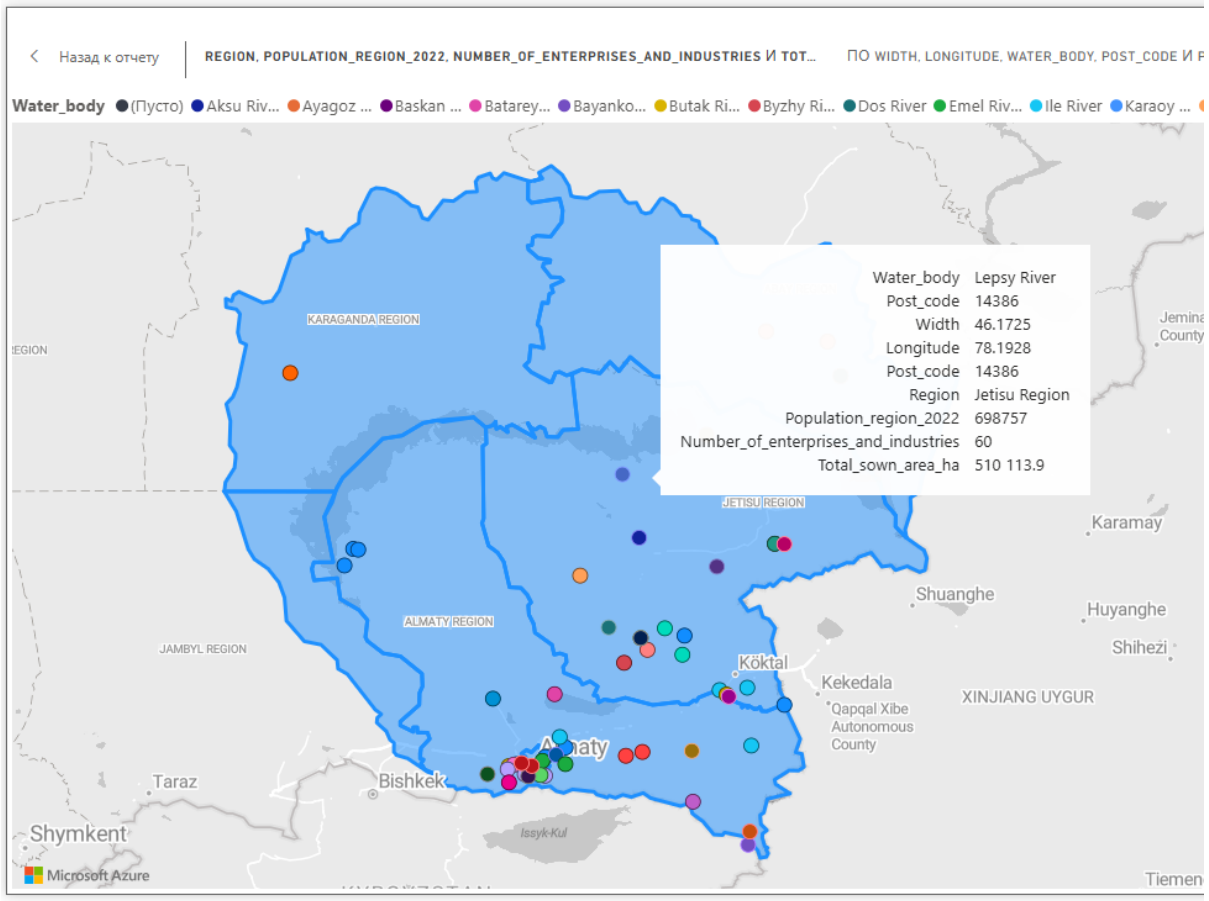


Рисунок 4.18 – Пространственная визуализация гидропостов и показателей

На рисунке 4.18 показана карта Иле-Балхашского бассейна с отображением гидропостов и соответствующих им атрибутивных данных. Географические координаты, сохранённые в формате WKT согласно стандарту GeoSPARQL, обеспечивают автоматическое отображение точек наблюдения.

При выборе объекта на карте отображаются:

- регион,
- численность населения,
- количество предприятий,
- измеренные концентрации,
- класс качества воды.

Таким образом, реализуется пространственно-аналитическая интеграция данных, позволяющая учитывать территориальные особенности при оценке экологической ситуации.

Chemical_Symbol	Regulated_indicators	Class_1_Max	Class_1_Min	Class_2_Max	Class_2_Min	Class_3_Max	Class_3_Min	Class_4_Max	Class_4_Min	Class_5_Max	Class_5_Min	Class_6_Max	Class_6_Min
Zntotal	Zinc (Total)	0.04	0.04	0.04	0.04	0.12	0.04	0.20	0.12	0.20	0.20		0.20
Zndissolved	Zinc (Dissolved)	0.01	0.01	0.01	0.01	0.03	0.01	0.05	0.03	0.05	0.05		0.05
Twater	Temperature												
TDS / Salinity	Total Mineralization / Salinity	1,000.00		1,300.00	1,000.00	1,500.00	1,300.00		1,500.00	2,000.00			2,000.00
SO42−	Sulfates	100.00		500.00	100.00		500.00	600.00		1,500.00			1,500.00
Si	Silicon	10.00	10.00	12.00	10.00	12.00	12.00	12.00	12.00	12.00	12.00		12.00
SCN−	Thiocyanates	0.10	0.10	0.10	0.10	0.15	0.10	0.20	0.15	0.20	0.20		0.20
Ptotal	Total Phosphorus, Ptotal	0.20	0.10	0.40	0.20	1.00	0.40	1.00	1.00	1.00	1.00	6.00	1.00
PO43−	Phosphates	0.40	0.20		0.40	1.00			1.00	3.50		6.00	3.50
PCl3	Phosphorus Trichloride	0.01	0.01	0.25	0.01	0.60	0.25	0.60	0.60	0.60	0.60	6.00	0.60
Pbtotal	Lead (Total)	0.60	0.12	0.60	0.60	1.00	0.60	1.00	1.00	1.00	1.00		1.00
Pbdissolved	Lead (Dissolved)	0.03	0.01	0.03	0.03	0.05	0.03	0.05	0.05	0.05	0.05		0.05
O2	Dissolved Oxygen		6.00	6.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.00	4.00
O2	Oxygen Saturation		0.90	0.90	0.80	0.80	0.60	0.60	0.40	0.40	0.60	0.20	0.40
Ntotal	Total Nitrogen, Ntotal	4.00	1.50	8.00	4.00	20.00	8.00	20.00	20.00	20.00	20.00		20.00
NO3−	Nitrate Ion	40.00		45.00	40.00	45.00	45.00	45.00	45.00	45.00	45.00		45.00
NO2−	Nitrite Ion	3.30	0.10	3.30	3.30	3.30	3.30	5.00	3.30	5.00	5.00		5.00
Nitotal	Nickel (Total)	0.02	0.01	0.05	0.03	0.10	0.05	0.10	0.10	0.10	0.10		0.10
Nidissolved	Nickel (Dissolved)	0.02	0.01	0.04	0.02	0.08	0.04	0.08	0.08	0.08	0.08		0.08
NH4+	Ammonium Ion	0.50		1.00	0.50	2.00	1.00	2.60	2.00	2.60	2.60		2.60
NH3-nH2O	Ammonia	0.05	0.05	0.10	0.05	0.20	0.10	0.26	0.20	0.26	0.26		0.26
Na+	Sodium	200.00	120.00	200.00	200.00	200.00	200.00	200.00	200.00	200.00	200.00		200.00
Mototal	Molybdenum (Total)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01		0.01
Modissolved	Molybdenum (Dissolved)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00
Mn2+	Manganese (II)	0.01	0.01	0.10	0.01	0.20	0.10	0.30	0.20	0.30	0.30		0.30
Mg2+	Magnesium	20.00		60.00	20.00		60.00	100.00		100.00	100.00		100.00
LPC	Lactose-Positive Coliform Bacteria, LPC	1,000.00	1,000.00	1,000.00	1,000.00	5,000.00	1,000.00	5,000.00	5,000.00		5,000.00		
K+	Potassium	50.00	50.00	50.00	50.00		50.00	100.00		100.00	100.00		100.00

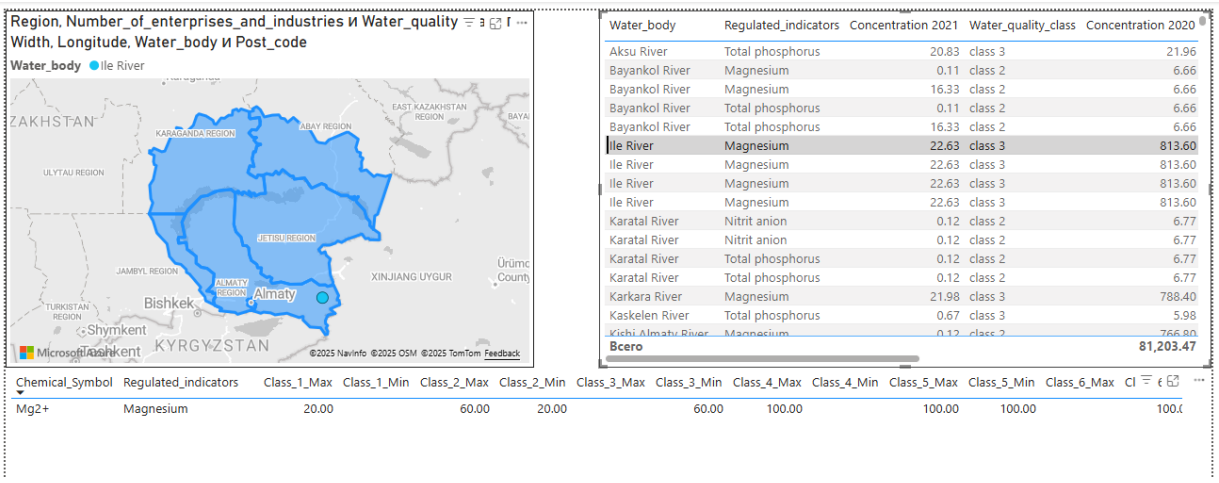
Рисунок 4.19 – Сопоставление нормативных диапазонов и фактических измерений

На рисунке 4.20 представлены нормативные диапазоны концентраций (Class_1_Min – Class_6_Max) и их сопоставление с фактическими измерениями. Это позволяет продемонстрировать воспроизводимость процедуры автоматической классификации.

Реализованный цикл обработки данных включает:

- поступление данных наблюдений,
- формирование RDF-триплетов,
- применение правил логического вывода,
- автоматическое присвоение класса качества,
- извлечение агрегированных данных через SPARQL,
- аналитическую и картографическую визуализацию.

Такой цикл подтверждает полноту архитектурного решения и его завершённость с точки зрения прикладного применения.



Water_body	Regulated_indicators	Concentration 2021	Water_quality_class	Concentration 2020
Aksu River	Total phosphorus	20.83	class 3	21.96
Bayankol River	Magnesium	0.11	class 2	6.66
Bayankol River	Magnesium	16.33	class 2	6.66
Bayankol River	Total phosphorus	0.11	class 2	6.66
Bayankol River	Total phosphorus	16.33	class 2	6.66
Ile River	Magnesium	22.63	class 3	813.60
Ile River	Magnesium	22.63	class 3	813.60
Ile River	Magnesium	22.63	class 3	813.60
Ile River	Magnesium	22.63	class 3	813.60
Karatal River	Nitrit anion	0.12	class 2	6.77
Karatal River	Nitrit anion	0.12	class 2	6.77
Karatal River	Total phosphorus	0.12	class 2	6.77
Karatal River	Total phosphorus	0.12	class 2	6.77
Karkara River	Magnesium	21.98	class 3	788.40
Kaskelen River	Total phosphorus	0.67	class 3	5.98
Kishi Almaty River	Magnesium	0.12	class 2	766.80
Всего				81,203.47

Chemical_Symbol	Regulated_indicators	Class_1_Max	Class_1_Min	Class_2_Max	Class_2_Min	Class_3_Max	Class_3_Min	Class_4_Max	Class_4_Min	Class_5_Max	Class_5_Min	Class_6_Max	Cl
Mg2+	Magnesium	20.00		60.00	20.00		60.00	100.00		100.00	100.00		100.(

Рисунок 4.20 – Интегрированная аналитическая панель мониторинга качества воды и региональных показателей

На рисунке 4.19 представлена аналитическая панель мониторинга, сформированная на основе данных, извлечённых из графовой базы посредством SPARQL-запросов и экспортированных в систему бизнес-аналитики. Данный интерфейс демонстрирует практическую реализацию интеграции экологических, гидрологических и социально-экономических параметров в едином аналитическом пространстве.

Левая часть панели содержит геопространственную визуализацию Иле-Балхашского бассейна с административным разграничением регионов. На карте отображаются водные объекты и соответствующие им точки мониторинга. Географические координаты, сохранённые в онтологии, обеспечивают корректное позиционирование объектов и возможность интерактивного выбора конкретной реки.

Правая верхняя часть панели представлена в виде таблицы, содержащей структурированные данные по каждому водному объекту. Таблица включает следующие атрибуты:

- наименование водного объекта (Water_body);
- регулируемый показатель (Regulated_indicators);
- концентрацию за 2021 год;
- автоматически определённый класс качества воды (Water_quality_class);
- концентрацию за 2020 год.

Наличие столбца *Water_quality_class* подтверждает, что отображаемые данные уже прошли процедуру автоматической классификации на основе нормативных диапазонов, заложенных в онтологической модели. Таким образом, пользователю предоставляется не только количественная информация, но и интерпретированный результат оценки состояния воды.

В нижней части панели отображаются нормативные диапазоны значений для соответствующего химического показателя (например, Mg^{2+} — Magnesium), включая минимальные и максимальные пороги для классов I–VI. Это позволяет сопоставлять фактическую концентрацию с установленными нормативными границами непосредственно в аналитическом интерфейсе.

Особую значимость представляет возможность одновременного анализа территориальных характеристик региона (численность населения, количество предприятий и др.) совместно с экологическими показателями. Такая интеграция создаёт основу для выявления взаимосвязей между промышленной нагрузкой и качеством водных ресурсов.

Таким образом, рисунок 4.19 демонстрирует завершённый цикл обработки данных: от сбора измерений и их формализации в графе знаний до аналитической визуализации и поддержки принятия решений. Панель подтверждает прикладную применимость разработанной модели и её готовность к использованию в системах экологического мониторинга регионального уровня.

Таким образом, представленный этап извлечения данных и их аналитической интерпретации подтверждает функциональную завершённость разработанной онтологически-ориентированной модели мониторинга водных ресурсов. Реализованный механизм SPARQL-доступа обеспечивает

формализованное объединение гидрологических наблюдений, нормативных порогов и социально-экономических характеристик в рамках единого информационного пространства, а последующая интеграция с аналитическими платформами демонстрирует практическую применимость системы. Полученные результаты свидетельствуют о том, что граф знаний выступает не только средством хранения и структурирования данных, но и основой для комплексного анализа, автоматической классификации качества воды и пространственно-ориентированного мониторинга. Тем самым подтверждается научная и прикладная значимость проделанной работы, обеспечивающей переход от разрозненных гетерогенных источников к целостной интеллектуальной системе поддержки принятия решений в области управления водными ресурсами Иле-Балхашского бассейна.

ЗАКЛЮЧЕНИЕ

В диссертационной работе разработана и реализована онтологически-ориентированная геоинформационная модель для извлечения, интеграции и логической интерпретации знаний из гетерогенных источников по водному домену Казахстана. Предложенная архитектура охватывает полный цикл обработки данных: нормализацию таблиц (PDF, Excel, веб-источники), итеративную семантическую интерпретацию (STI), маппинг в RDF, онтологическое выравнивание, загрузку в GraphDB и последующую аналитическую эксплуатацию.

Для объективной оценки качества метода сформирован эталонный набор триплетов на основе 11 864 записей и 62 атрибутов. Экспериментальная валидация показала высокую точность извлечения знаний (F1 до 1.000 по ключевым категориям) и устойчивость к гетерогенности данных. Сравнительный анализ с TableMiner++, QTLTableMiner++ и MantisTable подтвердил превосходство разработанного подхода: итоговый Triple F1 = 0.982, при Triple Precision = 0.965 и Triple Recall = 0.965. Относительный прирост по метрике сопоставления предикатов достигает 50% по сравнению с QTLTableMiner++, что подтверждает эффективность механизма семантического сопоставления и итеративной дезамбигуации.

Метод был применён к полному корпусу данных Иле-Балхашского бассейна (25 гетерогенных таблиц), в результате чего сформирован масштабируемый граф знаний объёмом 502 720 явных триплетов. Интеграция SWRL-правил привела к генерации 105 527 выводимых утверждений и увеличению общего объёма графа до 588 341 триплетов. Коэффициент расширения составил 1.22, что означает формирование в среднем 22 дополнительных логически обоснованных утверждений на каждые 100 исходных фактов. Это свидетельствует о переходе модели от декларативного хранилища данных к дедуктивной семантической системе.

Разработанная гидро-онтологическая модель выровнена с международными стандартами (SOSA/SSN, WHOW, HY_Features, Time Ontology) и дополнена национальными модулями Water_Regulations_KZ и Socio_Geo_KZ, что обеспечивает интероперабельность и формализацию нормативных требований Республики Казахстан.

Завершающим этапом стала реализация аналитической эксплуатации графа знаний с использованием SPARQL-запросов. Продемонстрирована возможность междисциплинарного извлечения данных (наблюдения – нормативы – регионы – социально-экономические параметры) и их экспорта в машиночитаемые форматы (CSV, JSON, TTL) для последующего использования в системах мониторинга и аналитических платформах.

Таким образом, поставленная цель диссертационного исследования достигнута. Создана методология и реализована масштабируемая онтологически-ориентированная геоинформационная модель водного домена, обеспечивающая интеграцию гетерогенных данных, автоматическое

формирование триплетов, логический вывод новых знаний и их аналитическую эксплуатацию. Полученные результаты подтверждают научную новизну, практическую значимость и перспективность разработанного подхода для задач мониторинга и управления водными ресурсами Казахстана.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] T. R. Gruber, “A translation approach to portable ontology specifications,” Knowledge Acquisition, vol. 5, no. 2, pp. 199–220, 1993.

[2] R. Studer, V. R. Benjamins, and D. Fensel, “Knowledge engineering: Principles and methods,” Data & Knowledge Engineering, vol. 25, no. 1–2, pp. 161–197, 1998.

[3] C. S. Jones et al., “Iowa statewide stream nitrate load calculated using in situ sensor network,” JAWRA Journal of the American Water Resources Association, vol. 54, no. 2, pp. 471–486, 2018.

[4] IDC, The Digitization of the World From Edge to Core, White Paper, 2018.

[5] I. Demir et al., “Data-enabled field experiment planning, management, and research using cyberinfrastructure,” Journal of Hydrometeorology, vol. 16, no. 3, pp. 1155–1170, 2015.

[6] M. Masmoudi, S. B. A. B. Lamine, H. B. Zghal, B. Archimede, and M. H. Karray, “Knowledge hypergraph-based approach for data integration and querying: Application to Earth Observation,” Future Generation Computer Systems, vol. 115, pp. 720–740, 2021.

[7] Y. Sermet and I. Demir, “Towards an information centric flood ontology for information management and communication,” Earth Science Informatics, vol. 12, no. 4, pp. 541–551, 2019.

[8] D. Mukhopadhyay and S. Shikalgar, “A model approach to build basic ontology,” arXiv preprint, arXiv:1311.6245, 2013.

[9] L. Ehrlinger and W. Wöß, “Towards a definition of knowledge graphs,” in Proc. 12th International Conference on Semantic Systems (SEMANTiCS 2016), pp. 1–4, 2016.

[10] A Haller, A., Janowicz, K., Cox, S., Lefrançois, M., Taylor, K., Phuoc, D., Lieberman, J., García-Castro, R., Atkinson, R., & Stadler, C. (2018). The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. Semantic Web, 10(1), 1–25. https://doi.org/10.3233/SW-180320.

[11] Wang X., Wei H., Chen N., He X., Tian Z. An Observational Process Ontology-Based Modeling Approach for River Water Quality Monitoring // Water. – 2020. – Vol. 12, No. 3. – Art. 715. – DOI: 10.3390/w12030715.

[12] Tripathi, A., & Babaie, H. A. (2008). Developing a modular hydrogeology ontology by extending the SWEET upper-level ontologies. Computers & Geosciences, 34(9), 1022–1033. https://doi.org/10.1016/j.cageo.2007.08.009.

[13] E. Katsiri, C. Makropoulos,An ontology framework for decentralized water management and analytics using wireless sensor networks,Desalination and Water Treatment,Volume 57, Issue 54,2016,Pages 26355-26368,ISSN 1944-3986,https://doi.org/10.1080/19443994.2016.1202144.

[14] Raskin, R. G., & Pan, M. J. (2005). Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences, 31(9), 1119–1125. https://doi.org/10.1016/j.cageo.2004.12.004.

[15] Sermet, Y. and Demir, I. (2018). An intelligent system on knowledge generation and communication about flooding. Environ. Modell. Softw. 108, 51-60. https://doi.org/10.1016/j.envsoft.2018.06.003.

[16] Hu, A. and Demir, I. (2021). Real-time flood mapping on client-side web systems using hand model. Hydrology. 8(2), 65. https://doi.org/10.1016/j.autcon.2017.02.004 .

[17] Yildirim, E. and Demir, I. (2021). An integrated flood risk assessment and mitigation framework: A case study for middle Cedar River Basin, Iowa, US. Int. J. Disaster Risk Reduct. 56, 102113. https:// doi.org/10.1016/j.ijdrr.2021.102113.

[18] NSTC (2018). Open Knowledge Network: Summary of the Federal NITRD Big Data IWG Workshop. National Science and Technology Council.

[19] Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Sci. Am. 284(5), 34-43. https://doi.org/10.1038/scientificamerican05 01-34.

[20] Chungoora, N., Young, R.I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N.A., Cutting-Decelle, A.F., Harding, J.A. and Case, K. (2013). A model-driven ontology approach for manufacturing system interoperability and knowledge sharing. Comput. Ind. 64(4), 392-401. https://doi.org/10.1016/j.compind.2013.01.003.
[21] Yoo, D. and No, S. (2014). Ontology-based economics knowledge sharing system. Expert Syst. Appl. 41(4), 1331-1341. https://doi.org/10.1016/j.eswa.2013.08.031
[22] Gayathri, R. and Uma, V. (2018). Ontology based knowledge representation technique, domain modeling languages and planners for robotic path planning: A survey. ICT Express. 4(2), 69-74. https: //doi.org/10.1016/j.icte.2018.04.008.
[23] Jelokhani-Niaraki, M. (2018). Knowledge sharing in Web-based collaborative multicriteria spatial decision analysis: An ontology-based multi-agent approach. Comput. Environ. Urban Syst. 72, 104-123. https://doi.org/10.1016/j.compenvurbsys.2018.05.012.
[24] Qi, J.D., Ding, L. and Lim, S. (2020). Ontology-based knowledge representation of urban heat island mitigation strategies. Sust. Cities Soc. 52, 101875. https://doi.org/10.1016/j.scs.2019.10187 5.
[25] Noy, N.F. and McGuinness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05SMI-2001-0880.
[26] Antoniou, G. and Harmelen, F.V. (2004). Web ontology language: Owl. Handbook on ontologies. Springer International Publishing, pp. 67-92.
[27] Middleton, S.E., Roure, D.D. and Shadbolt, N.R. (2004). Ontology-based recommender systems. Handbook on ontologies. Springer International Publishing, pp 477-498. https://doi.org/10.1007/97 8-3-540-92673-335.
[28] Kalibatiene, D. and Vasilecas, O. (2011). Survey on ontology lan- guages. International Conference on Business Informatics Re- search. Riga, 124-141.
[29] Slimani, T. (2015). Ontology development: A comparing study on tools, languages and formalisms. Indian. J. Sci. Technol. 8(24), 1-12. https://doi.org/10.17485/IJST/2015/V8I24/54249.
[30] Falconer, S. (2010). OntoGraf. https://protegewiki.stanford.edu/wiki/OntoGraf (Accessed April 12, 2023).
[31] Gil, Y., Ratnakar, V. and Garijo, D. (2015). OntoSoft: Capturing scien- tific software metadata. Proc. of the 8th International Conference on Knowledge Capture, New York, 1-4.
[32] Dudáš, M., Lohmann, S., Svátek, V. and Pavlov, D. (2018). Ontology visualization methods and tools: a survey of the state of the art. Knowl. Eng. Rev. 33, E10. https://doi.org/10.1017/S0269888918 000073.
[33] Open Geospatial Consortium (OGC). WaterML 2: Part 3 – Surface Hydrology Features (HY_Features). Version 1.0. OGC 14-111r6, 2018. Available: https://docs.ogc.org/is/14-111r6/14-111r6.html.
[34] Sinha, G., Mark, D.M., Kolas, D., Varanka, D., Romero, B.E., Feng, C.C., Usery, E.L., Liebermann, J. and Sorokine, A. (2014). An ontology design pattern for surface water features, International Conference on Geographic Information Science. Vienna, 187–203.
[35] European Commission. INSPIRE Ontology – OWL representations of INSPIRE conceptual models. Version 4.0. Luxembourg: Publications Office of the European Union, 2021. DOI: 10.28704/j.pangaea.ins.ontology.v4.
[36] Vilches-Blázquez, L.M., Ramos, J.A., López-Pellicer, F.J., Corcho, O. and Nogueras-Iso, J. (2009). An approach to comparing different ontologies in the context of hydrographical information. Information fusion and geographic information systems. Springer International Publishing, pp. 193–207.
[37] Varanka, D.E. and Cheatham, M. (2016). Spatial concepts for hydrography ontology alignment. AutoCarto 2016, New Mexico, 224–237.
[38] Bermudez, L.E. and Piasecki, M. (2003). Hydrologic ontology for the web. American Geophysical Union: Fall 2003 Conference, San Francisco, U41B-0016.
[39] Hahmann, T., Stephen, S. and Brodaric, B. (2016). Semantically refining the groundwater

markup language (GWML2) with the help of a reference ontology. International Conference on GIScience Short Paper Proceedings. Montreal, 1–1, 118–121.
[40] Baydaroğlu Ö., Güneş E.O., Taş E., Yıldırım E., Öztürk D., Çelik F. A Comprehensive Review of Ontologies in Hydrology: Classification, Comparison, and Application for Decision Support // Journal of Earth Science and Engineering. – 2023. – Vol. 13, No. 4. – P. 270–286.
[41] Durnford D.S., Leadbetter A., Cox S.J.D., Taylor P. OGC WaterML 2: Part 3 – Surface Hydrology Features (HY_Features) Conceptual Model. – Open Geospatial Consortium, 2018. – 99 p. – (OGC 15-111r1).
[42] Janowicz K., Haller A., Cox S.J.D., Phuoc D.L., Lefrançois M. SOSA: A Lightweight Ontology for Sensors, Observations, Samples, and Actuators // Journal of Web Semantics. – 2019. – Vol. 56. – Art. 100490. – DOI: 10.1016/j.websem.2018.06.003.
[43] Shen, G., Huang, Z., Zhu, X., Wang, L., Xiang, G.: Using description logics reasoner for ontology matching. In: Intelligent Information Technology Application, Workshop on. pp. 30–33. IEEE (2007).
[44] Furletti B., Lodi S., Martoglia R., Ronzano F. Knowledge Discovery in Ontologies // CEUR Workshop Proceedings. – 2012. – Vol. 858. – P. 1–12.
[45] Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al.: Swrl: A semantic web rule language combining owl and ruleml. W3C Member submission 21, 79 (2004).
[46] McGuinness, D.L., Van Harmelen, F., et al.: Owl web ontology language overview. W3C recommendation 10(10), 2004 (2004).
[47] Lawan A., Rakib A. The Semantic Web Rule Language Expressiveness Extensions – A Survey // arXiv preprint arXiv:1903.11723 [cs.AI]. – 2019. – DOI: 10.48550/arXiv.1903.11723.
[48] Horrocks, I., Patel-schneider, P.F., Bechhofer, S., Tsarkov, D.: OWL rules: A proposal and prototype implementation. Web Semantics: Science, Services and Agents on the World Wide Web 3(1), 2340 (jul 2005), http://www.sciencedirect.com/science/article/pii/\S1570826805000053.
[49] Rosati, R.: Semantic and computational advantages of the safe integration of ontologies and rules. Lect. Notes Comput. Sci. 3703 LNCS, 5064 (2005).
[50] OConnor, M., Das, A.: Sqwrl: A query language for owl. In: Proceedings of the 6th International Conference on OWL: Experiences and Directions - Volume 529. pp. 208–215. OWLED09, CEUR-WS.org, Aachen, Germany (2009), http://dl.acm.org/citation.cfm?id=2890046.2890072.
[51] Horn, A.: On Sentences Which are True of Direct Unions of Algebras. J. Symb. Log. 16(1), 14–21 (1951).
[52] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York, NY, USA (2003).
[53] Ahmedi, Lule & Jajaga, Edmond & Ahmedi, Figene. (2013). An Ontology Framework for Water Quality Management. CEUR Workshop Proceedings. 1063. https://ceur-ws.org/Vol-1063/paper3.pdf.
[54] Wang, P., Zheng, J.G., Fu, L., Patton, E.W., Lebo, T., Ding, L., Liu, Q., Luciano, J.S., McGuinness, D.L.: TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring. TWC RPI, Troy, NY, (2011).
[55] INWATERSENSE Core Ontology [Электронный ресурс]. – Режим доступа: http://inwatersense.uni-pr.edu/ontologies/inws-core.owl (дата обращения: 11.08.2025).
[56] INWATERSENSE Regulations Ontology [Электронный ресурс]. – Режим доступа: http://inwatersense.uni-pr.edu/ontologies/inws-regulations.owl (дата обращения: 11.08.2025).
[57] http://www.w3.org/TR/owl-time/
[58] http://www.w3.org/2003/01/geo/
[59] Directive 2000/60/EC of the European Parliament and of the Council of Europe of 23 October

2000 establishing a framework for Community action in the Field of water quality O.J. L327/1 (2000).
[60] Abdelmageed, N., Schindler, S. JenTab: Matching Tabular Data to Knowledge Graphs. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), pp. 40–49 (2020)
[61] Abdelmageed, N., Schindler, S. JenTab: A Toolkit for Semantic Table Annotations. In: 2nd International Workshop on Knowledge Graph Construction (2021)
[62] Abdelmageed, N., Schindler, S. JenTab Meets SemTab 2021's New Challenges. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2021)
[63] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S. DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics 7(3), 154–165 (2009)
[64] Vrandecic, D., Krötzsch, M. Wikidata: A Free Collaborative Knowledge Base. Communications of the ACM 57(10), 78–85 (2014)
[65] Ehrlinger, L., Wöß, W. Towards a Definition of Knowledge Graphs. SEMANTiCS 48(1–4), 2 (2016)
[66] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J.E.L., Kirrane, S., Neumaier, S., Polleres, A., et al. Knowledge Graphs. arXiv:2003.02320 (2020)
[67] Thornton, P.K., Stroud, A., Hatibu, N., Legg, C., Ly, S., Twomlow, S., Molapong, K., Notenbaert, A., Kruska, R., von Kaufmann, R. Site Selection to Test an Integrated Approach to Agricultural Research for Development: Combining Expert Knowledge and Participatory Geographic Information System Methods. International Journal of Agricultural Sustainability 4(1), 39–60 (2006)
[68] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. Journal of Biomedical Informatics 41(5), 706–716 (2008)
[69] Van Eeden, W., De Villiers, J.P., Berndt, R., Nel, W.A., Blasch, E. Micro-Doppler Radar Classification of Humans and Animals in an Operational Environment. Expert Systems with Applications 102, 111 (2018)
[70] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: ACM SIGMOD International Conference on Management of Data (2008)
[71] Tanon, T.P., Weikum, G., Suchanek, F. YAGO 4: A Reasonable Knowledge Base. In: European Semantic Web Conference (ESWC), pp. 583–596. Springer (2020)
[72] Bizer, C., Heath, T., Berners-Lee, T. Linked Data: The Story So Far. International Journal on Semantic Web and Information Systems, pp. 205–227. IGI Global (2009)
[73] Bhagavatula, C.S., Noraset, T., Downey, D. Methods for Exploring and Mining Tables on Wikipedia. In: ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, pp. 18–26 (2013)
[74] Cafarella, M.J., Halevy, A., Wang, D.Z., Wu, E., Zhang, Y. Webtables: Exploring the Power of Tables on the Web. VLDB Endowment, pp. 538–549 (2008)
[75] Pasupat, P., Liang, P. Compositional Semantic Parsing on Semi-Structured Tables. arXiv:1508.00305 (2015)
[76] Sun, H., Ma, H., He, X., Yih, W.t., Su, Y., Yan, X. Table Cell Search for Question Answering. In: 25th International Conference on World Wide Web, pp. 771–782 (2016)
[77] Yin, P., Neubig, G., Yih, W.t., Riedel, S. TaBERT: Pre-Training for Joint Understanding of Textual and Tabular Data (2020)
[78] Ritze, D., Lehmberg, O., Bizer, C. Matching HTML Tables to DBpedia. In: 5th International Conference on Web Intelligence, Mining and Semantics, pp. 1–6 (2015)
[79] Yakout, M., Ganjam, K., Chakrabarti, K., Chaudhuri, S. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In: ACM SIGMOD International Conference on Management of Data, pp. 97–108 (2012)
[80] Zhang, S., Meij, E., Balog, K., Reinanda, R. Novel Entity Discovery from Web Tables. In: The

Web Conference, pp. 1298–1308 (2020)
[81] Zhang, S., Balog, K. Recommending Related Tables. arXiv:1907.03595 (2019)
[82] Benjelloun, O., Chen, S., Noy, N. Google Dataset Search by the Numbers. In: International Semantic Web Conference (ISWC), In-Use Track (2020)
[83] Noy, N., Burgess, M., Brickley, D. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In: 28th The Web Conference (WWW) (2019)
[84] Zhang, S., Balog, K. Web Table Extraction, Retrieval, and Augmentation: A Survey. ACM Transactions on Intelligent Systems and Technology (TIST) 11(2), 135 (2020)
[85] Bontto, S., Casiraghi, E., Mesiti, M. Table Understanding Approaches for Extracting Knowledge from Heterogeneous Tables. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p. e1407 (2021)
[86] Badaro, G., Saeed, M., Papotti, P. Transformers for Tabular Data Representation: A Survey of Models and Applications. Tech. Rep., EURECOM (2021), https://www.eurecom.fr/publication/6721
[87] Borisov, V., Leemann, T., Seiler, K., Haug, J., Pawelczyk, M., Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. arXiv:2110.01889 (2021)
[88] Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I. Information Extraction Meets the Semantic Web: A Survey. Semantic Web Journal 11(2), 255–235 (2020)
[89] Limaye, G., Sarawagi, S., Chakrabarti, S. Annotating and Searching Web Tables Using Entities, Types and Relationships. Proceedings of the VLDB Endowment 3(1–2), 1338–1347 (2010)
[90] Lautert, L.R., Scheidt, M.M., Dorneles, C.F. Web Table Taxonomy and Formalization. ACM SIGMOD Record 42(3), 28–33 (2013)
[91] Crestan, E., Pantel, P. Web-scale Table Census and Classification. In: 4th ACM International Conference on Web Search and Data Mining (WSDM), pp. 545–554 (2011)
[92] Eberius, J., Braunschweig, K., Hentsch, M., Thiele, M., Ahmadov, A., Lehner, W. Building the Dresden Web Table Corpus: A Classification Approach. In: IEEE 2nd International Symposium on Big Data Computing (BDC), pp. 41–50. IEEE (2015)
[93] Lehmberg, O., Ritze, D., Meusel, R., Bizer, C. A Large Public Corpus of Web Tables Containing Time and Context Metadata. In: 25th International Conference Companion on World Wide Web, pp. 75–76 (2016)
[94] Ritze, D. Web-Scale Web Table to Knowledge Base Matching. Ph.D. thesis, University of Mannheim (2017)
[95] Wang, Z., Dong, H., Jia, R., Li, J., Fu, Z., Han, S., Zhang, D. TUTA: Tree-Based Transformers for Generally Structured Table Pre-Training. In: 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1780–1790 (2021)
[96] Jimenez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K. SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In: European Semantic Web Conference (ESWC), pp. 514–530. Springer (2020)
[97] Ritze, D., Bizer, C. Matching Web Tables to DBpedia–A Feature Utility Study. In: International Conference on Extending Database Technology (EDBT), pp. 210–221 (2017)
[98] Zhang, Z. Effective and Efficient Semantic Table Interpretation Using TableMiner+. Semantic Web 8(6), 921–957 (2017)
[99] Huynh, V.P., Liu, J., Chabot, Y., Labbe, T., Monnin, P., Troncy, R. DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2020)
[100] Morikawa, H. Semantic Table Interpretation Using LOD4ALL. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), pp. 49–56 (2019)
[101] Deng, X., Sun, H., Lees, A., Wu, Y., Yu, C. TURL: Table Understanding through Representation Learning. arXiv:2006.14806 (2020)
[102] Wang, D., Shiralkar, P., Lockard, C., Huang, B., Dong, X.L., Jiang, M. TCN: Table Convolutional Network for Web Table Interpretation. arXiv:2102.09460 (2021)
[103] Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, C., Chen, C., Tan, W.C. Annotating Columns

with Pre-trained Language Models. arXiv:2104.01785 (2021)
[104] Chen, S., Karaoglu, A., Negreanu, C., Ma, T., Yao, J.G., Williams, J., Gordon, A., Lin, C.Y. Linkingpark: An Integrated Approach for Semantic Table Interpretation. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2020)
[105] Venetis, P., Halevy, A.Y., Madhavan, J., Pasca, M., Shen, W., Wu, F., Miao, G. Recovering Semantics of Tables on the Web. PVLDB 4(9), 528–538 (2011)
[106] Wang, J., Wang, H., Wang, Z., Zhu, K.Q. Understanding Tables on the Web. In: International Conference on Conceptual Modeling, pp. 141–155. Springer (2012)
[107] Sekhavat, Y.A., Di Paolo, F., Barbosa, D., Merialdo, P. Knowledge Base Augmentation Using Tabular Data. In: LDOW (2014)
[108] Oliveira, D., d'Aquin, M. ADog: Annotating Data with Ontologies and Graphs. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2019)
[109] Thawani, A., Hu, M., Hu, E., Zafar, H., Divvala, N.T., Singh, A., Qasemi, E., Szekely, P.A., Pujara, J. Entity Linking to Knowledge Graphs to Infer Column Types and Properties. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), vol. 2019, pp. 25–32 (2019)
[110] Khurana, U., Galhotra, S. Semantic Annotation for Tabular Data (2019)
[111] Steenwinckel, B., De Turck, F., Ongeane, F. MAGIC: Mining an Augmented Graph Using INK, Starting from a CSV. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2021)
[112] Alobaid, A., Corcho, O. Balancing coverage and specificity for semantic labelling of subject columns. Knowledge-Based Systems, p. 108092 (2022)
[113] Zwicklbauer, S., Einsiedler, C., Granitzer, M., Seifert, C. Towards Disambiguating Web Tables. In: International Semantic Web Conference (Posters & Demos), pp. 205–208 (2013)
[114] Steenwinckel, B., Vandewiele, G., De Turck, F., Ongenae, F. Csv2KG: Transforming Tabular Data into Semantic Knowledge. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2019)
[115] Nguyen, P., Kertkeidkachorn, N., Ichise, R., Takeda, H. MTab: Matching Tabular Data to Knowledge Graph Using Probability Models. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2019)
[116] Nguyen, P., Yamada, I., Kertkeidkachorn, N., Ichise, R., Takeda, H. Mtab4Wikidata at SemTab 2020: Tabular Data Annotation with Wikidata. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2020)
[117] Nguyen, P., Yamada, I., Kertkeidkachorn, N., Ichise, R., Takeda, H. SemTab 2021: Tabular Data Annotation with MTab Tool. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2021)
[118] Huynh, V.P., Chabot, Y., Labbe, T., Liu, J., Troncy, R. From Heuristics to Language Models: A Journey Through the Universe of Semantic Table Interpretation with DAGOBAH. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2022)
[119] Huynh, V.P., Liu, J., Chabot, Y., Deuze, F., Labbe, T., Monnin, P., Troncy, R. DAGOBAH: Table and Graph Contexts for Efficient Semantic Annotation of Tabular Data. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2021)
[120] Cremaschi, M., Avogadro, R., Barazzetti, A., Chieregato, D. MantisTable SE: An Efficient Approach for the Semantic Table Interpretation. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2020)
[121] Cremaschi, M., Avogadro, R., Chieregato, D. MantisTable: An Automatic Approach for the Semantic Table Interpretation. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), pp. 15–24 (2019)
[122] Mulwad, V., Finin, T., Joshi, A. Semantic Message Passing for Generating Linked Data from Tables. In: 12th International Semantic Web Conference (ISWC), pp. 363–378. Springer (2013)
[123] Mulwad, V., Finin, T., Syed, Z., Joshi, A., et al. Using Linked Data to Interpret Tables. In: 1st International Workshop on Consuming Linked Data (COLD) (2010)

[124] Ramnandan, S.K., Mittal, A., Knoblock, C.A., Szekely, P. Assigning Semantic Labels to Data Sources. In: European Semantic Web Conference (ESWC), pp. 403–417. Springer (2015)
[125] Minh, P., Suresh, A., Craig, A.K., Pedro, S. Semantic Labeling: A Domain-Independent Approach. In: 15th International Semantic Web Conference (ISWC), pp. 446–462 (2016)
[126] Neumaier, S., Umbrich, J., Parreira, J.X., Polleres, A. Multi-Level Semantic Labelling of Numerical Values. In: 15th International Semantic Web Conference (ISWC), pp. 428–445 (2016)
[127] Kacprzak, E., Gimenez-Garcia, J.M., Piscopo, A., Koesten, L., Ibanez, L.D., Tennison, J., Simperl, E. Making Sense of Numerical Data–Semantic Labelling of Web Tables. In: European Knowledge Acquisition Workshop, pp. 163–178. Springer (2018)
[128] Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V. Matching Web Tables with Knowledge Base Entities: From Entity Lookups to Entity Embeddings. In: 16th International Semantic Web Conference (ISWC), pp. 260–277. Springer (2017)
[129] Biswas, R., Turker, R., Moghaddam, F.B., Koutraki, M., Sack, H. Wikipedia Infobox Type Prediction Using Embeddings. In: DL4KGS@ESWC, pp. 46–55 (2018)
[130] Chabot, Y., Labbe, T., Liu, J., Troncy, R. DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System. In: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, pp. 41–48 (2019)
[131] Liu, J., Huynh, V.P., Chabot, Y., Troncy, R. Radar Station: Using KG Embeddings for Semantic Table Interpretation and Entity Disambiguation. In: 21st International Semantic Web Conference (ISWC) (2022)
[132] Hulsebos, M., Hu, K., Bakker, M., Zgraggen, E., Satyanarayan, A., Kraska, T., Demiralp, C., Hidalgo, C. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. In: 25th ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD), pp. 1500–1508 (2019)
[133] Zhang, D., Suhara, Y., Li, J., Hulsebos, M., Demiralp, C., Tan, W.C. Sato: Contextual Semantic Type Detection in Tables (2019)
[134] Chen, J., Jimenez-Ruiz, E., Horrocks, I., Sutton, C. ColNet: Embedding the Semantics of Web Tables for Column Type Prediction. In: 33rd AAAI International Conference on Artificial Intelligence (2018)
[135] Guo, T., Shen, D., Nie, T., Kou, Y. Web Table Column Type Detection Using Deep Learning and Probability Graph Model. In: International Conference on Web Information Systems and Applications, pp. 401–414. Springer (2020)
[136] Singh, G., Singh, S., Wong, J., Saari, A. Relation Extraction from Tables Using Artificially Generated Metadata. arXiv:2108.10750 (2021)
[137] Zhou, Y., Singh, S., Christodoulopoulos, C. Tabular Data Concept Type Detection Using Star-Transformers. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 3677–3681 (2021)
[138] Banko, M., Etzioni, O. The Trade-offs Between Open and Traditional Relation Extraction. In: ACL-08: HLT, pp. 28–36 (2008)
[139] Zhang, Z. Towards Efficient and Effective Semantic Table Interpretation. In: 3rd International Semantic Web Conference, pp. 487–502. Springer (2014)
[140] Robertson, S., Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. Now Publishers Inc (2009)
[141] Nguyen, P., Nguyen, K., Ichise, R., Takeda, H. EmbNum: Semantic Labeling for Numerical Values with Deep Metric Learning. In: Joint International Semantic Technology Conference, pp. 119–135. Springer (2018)
[142] Cremaschi, M., De Paoli, F., Rula, A., Spahiu, B. A Fully Automated Approach to a Complete Semantic Table Interpretation. Future Generation Computer Systems 112, 478–500 (2020)
[143] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. Advances in Neural Information Processing Systems 26 (2013)

[144] Wang, Z., Zhang, J., Feng, J., Chen, Z. Knowledge Graph Embedding by Translating on Hyperplanes. In: AAAI Conference on Artificial Intelligence (2014)
[145] Ristoski, P., Paulheim, H. RDF2Vec: RDF Graph Embeddings for Data Mining. In: International Semantic Web Conference, pp. 498–514. Springer (2016)
[146] Guo, J., Fan, Y., Ai, Q., Croft, W.B. A Deep Relevance Matching Model for Ad-hoc Retrieval. In: 25th ACM International Conference on Information and Knowledge Management, pp. 55–64 (2016)
[147] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. Attention Is All You Need. arXiv:1706.03762 (2017)
[148] Cremaschi, M., Siano, A., Avogadro, R., Jimenez-Ruiz, E., Maurino, A. STILTool: A Semantic Table Interpretation Evaluation Tool. In: European Semantic Web Conference, pp. 61–66. Springer (2020)
[149] РГП «Казгидромет». Бассейны рек оз. Балхаш и оз. Алаколь (2002–2023) [Электронный ресурс] // Официальный сайт РГП «Казгидромет». – Режим доступа: https://kazhydromet.kz/ru/gidrologiya/basseyny-rek-oz-balkash-i-alakol-2002-2023 (дата обращения: 13.08.2025). – Загл. с экрана
[150] Или-Балхашский бассейн [Электронный ресурс] // Википедия, свободная энциклопедия. – Режим доступа: https://ru.wikipedia.org/wiki/Или-Балхашский_бассейн (дата обращения: 13.08.2025).
[151] Lake Balkhash (англ.). – Информация о водоёме на сайте базы данных по озёрам мира Международного комитета по проблемам озёр (International Lake Environment Committee World Lake Database, ILEC WLDB): wldb.ilec.or.jp (англ.). Дата обращения: 21 сентября 2022.
[152] Гидрологическая база данных РК. (2025). *ecodata.kz: Hydrological Data Portal*. Получено 25 октября 2025, с http://ecodata.kz:3838/app_hydro/
[153] Министерство водных ресурсов и ирригации Республики Казахстан. Об утверждении единой системы классификации качества воды в поверхностных водных объектах и (или) их частях: приказ от 04.06.2025 № 111-НҚ [Электронный ресурс] // Информационно-правовая система нормативных правовых актов Республики Казахстан «Әділет». – Режим доступа: https://adilet.zan.kz/rus/docs/G25MA000111 (дата обращения: 14.08.2025). – Загл. с экрана.
[154] Ospan A, Mansurova M, Kakimzhanov E, Ixanov S, Barakhnin V. Development of a Program for the Integration of Socio-Economic Indicators with Spatial Data to Analyze the Standard of Living of the Population of Kazakhstan. Vestn Nats Inzh Akad Respub Kaz. 2022;(85):67-78. doi:10.47533/2020.1606-146X.170.
[155] РГП «Казгидромет». Ежемесячный информационный бюллетень о состоянии окружающей среды [Электронный ресурс] // Раздел «Экология». – Режим доступа: https://kazhydromet.kz/ru/ecology/ezhemesyachnyy-informacionnyy-byulleten-o-sostoyanii-okruzhayuschey-sredy (дата обращения: 14.08.2025). – Загл. с экрана
[156] PyMuPDF Developers. (2024). PyMuPDF documentation (v1.24). Read the Docs. https://pymupdf.readthedocs.io/
[157] Artifex Software, Inc. (2024). *MuPDF documentation*. https://mupdf.com/docs/
[158] Оспан Ә., Мансурова М. Разработка гибридной модели для эффективного распределения водных ресурсов на основе модели прогнозирования. «Физика-математика ғылымдары» сериясы серия «физико-математические науки» series of physics & mathematical sciences, 2021 №1(69)
[159] Nugumanova A, Apayev K, Baiburin Y, Mansurova M, Ospan A. QURMA: A Table Extraction Pipeline for Knowledge Base Population. J Math M ech Comput Sci. 2022;114(2). https://bm.kaznu.kz/index.php/kaznu/article/view/1086. Published June 2022. Accessed October 19, 2022. doi:10.26577/JMMCS.2022.v114.i2.08.
[160] Cafarella, M. J., Halevy, A., Wang, D. Z., Wu, E., & Zhang, Y. (2008). WebTables: Exploring the power of tables on the web. *Proceedings of the VLDB Endowment, 1*(1), 538–549. vldb.org,

Gatterbauer, W., Bohunsky, P., Herzog, M., Krüpl, B., & Pollak, B. (2007). Towards domain-independent information extraction from web tables. *Proceedings of the 16th International Conference on World Wide Web* (WWW). (Обсуждает восстановление структуры и семантики HTML-таблиц). vldb.org

[161] Beautiful Soup. (n.d.). *Beautiful Soup documentation*. Retrieved from the official bs4 documentation.. dl.acm.org

[162] Lehmberg, O., Ritze, D., Meusel, R., & Bizer, C. (2016). A large public corpus of web tables containing time and context metadata. *Proceedings of the 25th International Conference on World Wide Web Companion* (WWW Companion), 75–76. (WDC Web Table Corpora). webdatacommons.org

[163] Barakhnin V, Mansurova M, Grigorieva I, Kozhemyakina O, Ospan A. TableProcessor: The Tool for the Analysis and the Interpretation of Web Tables to Create the Geo Knowledge Base of Kazakhstan. In: Dolinina O, et al., eds. Artificial Intelligence in Models, Methods and Applications. AIES 2022. Studies in Systems, Decision and Control, vol 457. Springer; 2023:219-229. doi:10.1007/978-3-031-22938-1_15. Accessed April 25, 2023

[164] Singh G., Kuzniar A., van Mulligen E. M., Gavai A., Bachem C. W., Visser R. G. F., Finkers R. QTLTableMiner++: semantic mining of QTL tables in scientific articles // BMC Bioinformatics. – 2018. – Vol. 19. – Article number: 183. – DOI: 10.1186/s12859-018-2165-7.

[165] Ospan A, Mansurova M, Barakhnin V, Nugumanova A, Titkov R. The Development of a Water Resource Monitoring Ontology as a Research Tool for Sustainable Regional Development. Data. 2023;8(11):162. doi:10.3390/data8110162.

[166] Mansurova M, Barakhnin V, Ospan A, Titkov R. Ontology-Driven Semantic Analysis of Tabular Data: An Iterative Approach with Advanced Entity Recognition. Appl Sci. 2023;13(19):10918. doi:10.3390/app131910918.

ПРИЛОЖЕНИЕ А – Авторское свидетельство



ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

СВИДЕТЕЛЬСТВО

О ВНЕСЕНИИ СВЕДЕНИЙ В ГОСУДАРСТВЕННЫЙ РЕЕСТР ПРАВ НА ОБЪЕКТЫ, ОХРАНЯЕМЫЕ АВТОРСКИМ ПРАВОМ

№ 39296 от «28» сентября 2023 года

Фамилия, имя, отчество, (если оно указано в документе, удостоверяющем личность) автора (ов):
Оспан Әсел Ғалымжанқызы, Мансурова Мадина Есимхановна

Вид объекта авторского права: **программа для ЭВМ**

Название объекта: **Итеративный алгоритм для семантического анализа таблиц из гетерогенных источников для пополнения графов знаний**

Дата создания объекта: **01.09.2023**

Құжат түпнұсқалығын http://www.kazpatent.kz/ru сайтының "Авторлық құқық" бөлімінде тексеруге болады. https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

Подписано ЭЦП

Е. Оспанов

ПРИЛОЖЕНИЕ Б – Список публикаций

Scopus

1. Ospan A, Mansurova M, Barakhnin V, Nugumanova A, Titkov R. The Development of a Water Resource Monitoring Ontology as a Research Tool for Sustainable Regional Development. Data. **2023**;8(11):162. doi:10.3390/data8110162. (**Q2 64%**)
2. Mansurova M, Barakhnin V, **Ospan A**, Titkov R. Ontology-Driven Semantic Analysis of Tabular Data: An Iterative Approach with Advanced Entity Recognition. Appl Sci. **2023**;13(19):10918. doi:10.3390/app131910918. (**Q1 79%)**

Международные конференции

1. Barakhnin V, Mansurova M, Grigorieva I, Kozhemyakina O, **Ospan A.** TableProcessor: The Tool for the Analysis and the Interpretation of Web Tables to Create the Geo Knowledge Base of Kazakhstan. In: Dolinina O, et al., eds. Artificial Intelligence in Models, Methods and Applications. AIES 2022. Studies in Systems, Decision and Control, vol 457. Springer; 2023:219-229. doi:10.1007/978-3-031-22938-1_15. Accessed April 25, **2023**.

Статьи, опубликованные в КОКСОН РК:

1. Nugumanova A, Apayev K, Baiburin Y, Mansurova M, **Ospan A**. QURMA: A Table Extraction Pipeline for Knowledge Base Population. J Math Mech Comput Sci. 2022;114(2). https://bm.kaznu.kz/index.php/kaznu/article/view/1086. Published June 2022. Accessed October 19, **2022**. doi:10.26577/JMMCS.2022.v114.i2.08.
2. **Ospan A**, Mansurova M, Kakimzhanov E, Ixanov S, Barakhnin V. Development of a Program for the Integration of Socio-Economic Indicators with Spatial Data to Analyze the Standard of Living of the Population of Kazakhstan. Vestn Nats Inzh Akad Respub Kaz. **2022**;(85):67-78. doi:10.47533/2020.1606-146X.170.
3. **Оспан Ә**., Мансурова М. Разработка гибридной модели для эффективного распределения водных ресурсов на основе модели прогнозирования. «Физика-математика ғылымдары» сериясы серия «физико-математические науки» series of physics & mathematical sciences, **2021** №1(69)

Авторское свидетельство:

Оспан Әсел, Мансурова Мадина Есимхановна. Итеративный алгоритм для семантического анализа таблиц из гетерогенных источников для пополнения графов знаний. Свидетельство о государственной регистрации программы для ЭВМ № 39296 от «28» сентября 2023 года.