АННОТАЦИЯ

к диссертации Әсел Оспан Ғалымжанқызы на соискание степени доктора философии (PhD) по образовательной программе «8D06105 – Системная инженерия» на тему: «Разработка геоинформационных моделей с использованием онтологического подхода для извлечения знаний из гетерогенных источников»

Актуальность: Актуальность данного исследования обусловлена отсутствием в Республике Казахстан единой интегрированной базы данных, обеспечивающей формирование целостной и системно согласованной картины состояния водных ресурсов страны. В настоящее время сведения о гидрологических характеристиках, химических показателях качества воды, результатах мониторинга, нормативных требованиях и социально-экономических параметрах регионов распределены по различным ведомственным источникам и представлены в разнородных форматах. Отсутствие концептуально согласованной структуры данных затрудняют комплексный анализ, сопоставление показателей и построение интегрированных моделей водного баланса и экологического состояния бассейнов.

Существенной проблемой является сложность выявления связей между элементами табличных данных при отсутствии их семантической интерпретации. Традиционные методы обработки ориентированы преимущественно на синтаксический анализ и не формализуют смысловые зависимости между сущностями, вследствие чего данные остаются изолированными, а значимые взаимосвязи между показателями качества воды, характеристиками водных объектов, нормативами и социально-экономическими параметрами регионов не выявляются автоматически. Это ограничивает построение причинно-следственных моделей и снижает обоснованность управленческих решений.

В этих условиях возникает необходимость формирования единого графа знаний по водным ресурсам Казахстана на основе онтологического подхода, обеспечивающего интеграцию гетерогенных источников. Такой граф структурирует предметную область в виде формализованных сущностей и отношений, обеспечивает их машиночитаемость и логическую связность и создаёт основу для разработки интеллектуальных геоинформационных приложений и поддержки научно обоснованных решений в сфере водной безопасности.

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).

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».

Объекты и предметы исследования.

– Данные о поверхностных водах Иле-Балхашского бассейна, представленные в гетерогенных источниках (мониторинговые таблицы, нормативные документы, статистические отчёты, геопространственные данные).

– Методы семантической интерпретации табличных данных и онтологического моделирования, обеспечивающие формирование и расширение графа знаний водного домена на основе гетерогенных источников.

Методы исследования:

– методы онтологического моделирования и онтологического выравнивания с международными стандартами;

– методы семантической интерпретации табличных данных, включая разработанный итеративный подход STI;

– методы правило-ориентированного логического вывода в графах знаний на основе SWRL;

– методы формализации пространственно-временных данных в рамках стандартов Semantic Web.

Теоретическая значимость работы. Формализация итеративного процесса STI как последовательности обнаружения ролей столбцов, выявления предикатных кандидатов и атрибутных колонок с последующим онтологическим выравниванием и выводом знаний на основе правил.

Практическая значимость работы. Создание воспроизводимого конвейера «Таблица → RDF → GraphDB → SPARQL → Визуализация», который автоматически пополняет граф знаний о водных ресурсах из PDF/Веб/Excel таблиц и обеспечивает интероперабельность данных по стандартам RDF.

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.

Список опубликованных работ. Результаты диссертационного исследования опубликованы в 6 работах. Из них 2 статьи в международных научных изданиях, входящих в базу данных Web of science и Scopus, 1 работа в материалах международных и республиканских конференций, 3 статьи в журналах, рекомендованных Комитетом по контролю в сфере образования и науки МОН РК, 1 авторское свидетельство по результатам практического внедрения диссертационных исследований:

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%).
3. 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.
4. 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.
5. 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.
6. 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)
7. 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.