ABSTRACT

for the dissertation of Assel Ospan
submitted for the degree of Doctor of Philosophy (PhD) in the educational program "8D06105 – Systems Engineering" on the topic: "Development of geoinformation models using an ontological approach to extract knowledge from heterogeneous sources"

Relevance of the research topic. Актуальность: The relevance of this study is обусловлена the absence in the Republic of Kazakhstan of a unified integrated database capable of providing a comprehensive and systemically consistent representation of the state of the country's water resources. Currently, information on hydrological characteristics, chemical water quality indicators, monitoring results, regulatory requirements, and socio-economic parameters of regions is distributed across various ведомственные sources and presented in heterogeneous formats. The lack of a conceptually consistent data structure complicates comprehensive analysis, comparison of indicators, and the development of integrated models of water balance and ecological conditions of river basins.

A significant problem is the difficulty of identifying relationships between elements of tabular data in the absence of their semantic interpretation. Traditional processing methods are mainly oriented toward syntactic analysis and do not formalize meaningful dependencies between entities. As a result, data remain isolated, and significant interrelations between water quality indicators, characteristics of water bodies, regulatory thresholds, and socio-economic parameters of regions are not automatically identified. This limits the construction of cause-and-effect models and reduces the validity of managerial decisions.

Under these conditions, there is a need to develop a unified knowledge graph for the water resources of Kazakhstan based on an ontological approach that ensures the integration of heterogeneous sources. Such a graph structures the domain in terms of formalized entities and relationships, ensures machine readability and logical consistency, and provides a foundation for the development of intelligent geoinformation applications and support of scientifically grounded decisions in the field of water security.

Level of research development. The interpretation of tabular data has been actively developing within the fields of Semantic Web and Knowledge Graph Engineering. Existing studies include heuristic methods (TableMiner, TableMiner++), hybrid and trainable pipelines (QTLTableMiner++, MantisTable, JenTab, MTab), as well as approaches based on deep models and knowledge graph representations (TURL, DODUO), addressing tasks of cell annotation, column typing, and relationship identification.

Despite this progress, most approaches are limited to mapping tables into RDF and do not cover the full data processing cycle, including ontological alignment, integration of heterogeneous sources, formalization of regulatory norms, and rule-based reasoning. In the water domain, international ontologies are not sufficiently adapted to national regulatory requirements and socio-geographical

contexts.

Issues related to the robustness of STI systems to heterogeneous sources and the formalization of regulatory computations in machine-readable form remain unresolved. Therefore, the development of an ontology-driven geoinformation model of the water domain, ensuring data integration and automated knowledge inference, remains a relevant scientific task.

Objective of the dissertation. To develop an ontology-driven geoinformation model for establishing relationships between objects of the water domain of Kazakhstan based on data from heterogeneous sources.

Research objectives:

1. To design the architecture of a geoinformation model based on an ontological approach for knowledge extraction.
2. To develop an iterative approach for semantic interpretation of tables with enhanced entity recognition.
3. To construct a hydro-ontological model aligned with international ontologies and extensible through additional modules.
4. To extend the ontology semantics and automate knowledge inference based on Semantic Web Rule Language (SWRL) rules.

Scientific novelty:

1. An improved iterative Semantic Table Interpretation (STI) approach is proposed, which outperforms existing methods (TableMiner++, QTLTableMiner++, MantisTable) through:

1.1. the application of enhanced heuristic rules for more accurate detection of subject columns and predicates;

1.2. the implementation of an automatic knowledge graph enrichment and updating mechanism based on extracted RDF triples.

2. A new hydro-ontological model of Kazakhstan has been developed, aligned with international ontologies (WHOW, HY_Features) and extended with:

2.1. a Water_Regulations_KZ module for water classification according to the Water Code of the Republic of Kazakhstan;

2.2. a Socio_Geo_KZ module linking water objects with socio-economic indicators of regions.

Main provisions submitted for defense:

1. An iterative Semantic Table Interpretation (STI) approach with enhanced heuristics for subject column and predicate detection, providing more accurate mapping of tables into RDF and outperforming TableMiner++, QTLTableMiner++, and MantisTable in terms of F1 metrics, with an embedded mechanism for automatic enrichment and updating of knowledge graphs.
2. A hydro-ontological model of Kazakhstan aligned with international standards (WHOW, HY_Features) and extended with the Water_Regulations_KZ and Socio_Geo_KZ modules, supporting rule-based reasoning (SWRL) for automatic assignment of water quality classes and identification of relationships "water bodies ↔ regulations ↔ observations ↔ regions."

Research objects and subjects.

– Data on surface waters of the Ile-Balkhash basin presented in heterogeneous sources (monitoring tables, regulatory documents, statistical reports, geospatial data).

– Methods of semantic interpretation of tabular data and ontology modeling enabling the formation and expansion of a water-domain knowledge graph based on heterogeneous sources.

Research methods:

– Ontology modeling and ontology alignment methods with international standards;

– Methods of semantic interpretation of tabular data, including the developed iterative STI approach;

– Rule-based logical reasoning methods in knowledge graphs based on SWRL;

– Methods for formalizing spatio-temporal data within the Semantic Web standards framework..

Theoretical significance of the study. The theoretical significance of the study lies in the formalization of the iterative STI process as a structured sequence of column role detection, identification of predicate candidates and attribute columns, followed by ontology alignment and rule-based knowledge inference.

Practical significance of the study. The practical significance consists in the development of a reproducible pipeline "Table → RDF → GraphDB → SPARQL → Visualization," which automatically enriches the water resources knowledge graph from PDF/Web/Excel tables and ensures data interoperability in accordance with RDF standards.

Reliability and validity of the research results. The reliability of the obtained results is ensured by the use of formalized Semantic Web standards, alignment of the developed ontological model with international ontologies, and experimental comparison of the proposed STI approach with existing methods, demonstrating its superiority in quality metrics. The validity of the results is confirmed through the application of the model to real data from the Ile-Balkhash basin, the logical consistency of the constructed knowledge graph, and the publication of the main research findings in peer-reviewed scientific journals.

Relationship of the topic to research programs. The dissertation research was carried out within the framework of research projects funded by the Ministry of Education and Science of the Republic of Kazakhstan (Science Committee of MES RK):

– Grant funding project AP09261344 "Development of Methods for Automatic Extraction of Geospatial Objects from Heterogeneous Sources for Information Support of Geoinformation Systems" (2021–2023);
– Program-targeted funding project BR24993001 "Development of a Large Language Model (LLM) to Support the Kazakh Language and Technological Progress" (2024–2026).

Approbation of the research. The main results of the dissertation were presented and discussed at international scientific conferences and academic seminars:

– AIES 2022: International Conference on Artificial Intelligence in Engineering and Science (Switzerland, 2022);

– Scientific seminar of the Department of Computer Science, Faculty of Information Technology, Al-Farabi Kazakh National University.

Structure and brief overview of the dissertation. The dissertation consists of an introduction, four chapters, a conclusion, and a list of references.

The introduction substantiates the relevance of the research topic, formulates the aim and objectives of the study, defines the object, subject, and research methods, and outlines the scientific novelty as well as the theoretical and practical significance of the work.

The first chapter presents the theoretical foundations of the ontological approach to water resource monitoring, analyzes existing ontologies in hydrology, and examines methods of semantic reasoning and knowledge extraction in hydrological models.

The second chapter investigates approaches to the semantic interpretation of tabular data, methods for evaluating the performance of STI systems, conducts the collection and analysis of heterogeneous data sources in the water domain of Kazakhstan, and reviews tools for extracting tables from unstructured sources.

The third chapter develops the methodology for constructing an ontology-driven geoinformation model, presents its multi-level and modular architecture, describes modules for observations, hydrographic objects, and national regulations, and introduces the iterative STI method along with a rule-based reasoning mechanism based on SWRL.

The fourth chapter provides the results of experimental validation of the proposed approach, presents a comparative analysis with existing semantic table interpretation methods, reports the formation of new axioms after applying SWRL rules, and demonstrates the implementation of an integrated hydro-social ontology in the GraphDB environment.

The conclusion summarizes the main results and findings of the dissertation research.

Publications based on the research results. The results of the dissertation research have been published in 6 scientific works, including 2 articles in international journals indexed in Web of Science and Scopus, 1 paper in the proceedings of international and national conferences, 3 articles in journals recommended by the Committee for Quality Assurance in Education and Science of MES RK, and 1 copyright certificate obtained as a result of the practical implementation of the dissertation research:

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. Ospan A., Mansurova M. Development of a Hybrid Model for Efficient Water Resource Allocation Based on a Forecasting Model. *Physics & Mathematical Sciences Series*, 2021, No. 1(69). (In Russian)

7. Copyright Certificate: Ospan Assel, Mansurova Madina Yesimkhanovna. Iterative Algorithm for Semantic Analysis of Tables from Heterogeneous Sources for Knowledge Graph Enrichment. Certificate of State Registration of a Computer Program No. 39296 dated September 28, 2023.