Әсел Оспан Ғалымжанқызының «8D06105 – Жүйелік инженерия» білім беру бағдарламасы бойынша философия докторы (PhD) дәрежесін алу үшін «Гетерогенді дереккөздерден білім алу үшін онтологиялық тәсілді қолдана отырып геоақпараттық модельдерді құру» тақырыбындағы диссертациялық жұмысына

АҢДАТПА

Зерттеу тақырыбының өзектілігі: Бұл зерттеудің өзектілігі Қазақстан Республикасында елдің су ресурстарының жай-күйі туралы кешенді және жүйелі түрде бірізді көрініс беретін бірыңғай интеграцияланған дерекқордың болмауынан туындайды. Қазіргі уақытта гидрологиялық сипаттамалар, судың химиялық сапасының көрсеткіштері, мониторинг нәтижелері, реттеуші талаптар және аймақтық әлеуметтік-экономикалық параметрлер туралы ақпарат әртүрлі ведомстволық дереккөздер бойынша шашыраңқы және әртүрлі форматтарда ұсынылған. Бөлшектелген сақтау және тұжырымдамалық тұрғыдан бірізді деректер құрылымының болмауы кешенді талдауға, көрсеткіштерді салыстыруға және бассейндердің су балансы мен экологиялық жағдайының интеграцияланған модельдерін құруға кедергі келтіреді.

Кестелік деректер элементтері арасындағы байланыстарды семантикалық интерпретациясыз анықтау күрделі мәселе болып табылады. Дәстүрлі өңдеу әдістері негізінен синтаксистік талдауға бағытталған және мәндік тәуелділіктерді формализацияламайды, соның салдарынан деректер оқшау күйде қалады, ал су сапасының көрсеткіштері, су объектілерінің сипаттамалары, нормативтік шектер мен аймақтардың әлеуметтік-экономикалық параметрлері арасындағы маңызды өзара байланыстар автоматты түрде анықталмайды. Бұл себеп-салдарлық модельдерді құруды шектеп, басқарушылық шешімдердің негізділігін төмендетеді.

Осы жағдайда гетерогенді дереккөздерді біріктіретін онтологиялық тәсілге негізделген Қазақстанның су ресурстары бойынша бірыңғай білім графын қалыптастыру қажеттілігі туындайды. Мұндай граф пәндік саланы формалданған сущностер мен қатынастар түрінде құрылымдайды, олардың машинамен оқылуын және логикалық байланыстылығын қамтамасыз етеді, сондай-ақ су қауіпсіздігі саласында ғылыми негізделген шешімдерді қолдайтын зияткерлік геоақпараттық қосымшаларды әзірлеуге негіз болады.

Мәселенің зерттелу деңгейі. Кестелік деректерді интерпретациялау мәселесі Semantic Web және Knowledge Graph Engineering бағыттары аясында белсенді дамуда. Қолданыстағы зерттеулер эвристикалық әдістерді (TableMiner, TableMiner++), гибридті және үйретілетін конвейерлерді (QTLTableMiner++, MantisTable, JenTab, MTab), сондай-ақ терең модельдер мен білім графтарының ұсынуларына негізделген тәсілдерді (TURL, DODUO) қамтиды, олар ұяшықтарды, баған типтерін және олардың арасындағы байланыстарды аннотациялау міндеттерін шешеді.

Дегенмен, көптеген тәсілдер кестелерді RDF-ке маппинг жасау

міндеттерімен шектеліп, онтологиялық сәйкестендіруді, гетерогенді дереккөздерді интеграциялауды, нормативтерді формализациялауды және ереже-бағытталған логикалық қорытындылауды қамтитын толық өңдеу циклін жүзеге асырмайды. Су ресурстары саласында халықаралық онтологиялар ұлттық нормативтік талаптар мен әлеуметтік-географиялық контекстке жеткілікті деңгейде бейімделмеген.

STI-жүйелерінің гетерогенді дереккөздерге тұрақтылығы және нормативтік есептеулерді машинамен оқылатын түрде формализациялау мәселелері әлі де толық шешімін таппаған. Осыған байланысты деректерді біріктіруді және жаңа білімді автоматты түрде шығаруды қамтамасыз ететін онтологиялық-бағытталған геоақпараттық модельді құру өзекті ғылыми міндет болып қала береді.

Диссертациялық жұмыстың мақсаты. Гетерогенді дереккөздер негізінде Қазақстанның су домені объектілері арасындағы байланыстарды анықтауға арналған онтологиялық-бағытталған геоақпараттық модельді әзірлеу.

Зерттеу міндеттері:

1. Білім алуға арналған онтологиялық тәсіл негізінде геоақпараттық модельдің архитектурасын жобалау.
2. Кеңейтілген мәндерді тану механизмі бар кестелерді семантикалық интерпретациялаудың итеративті тәсілін әзірлеу.
3. Халықаралық онтологиялармен сәйкестендіруді қамтамасыз ететін және жаңа модульдермен кеңейтілетін гидро-онтологиялық модельді құру.
4. Онтологияның семантикасын кеңейту және Semantic Web Rule Language (SWRL) ережелері негізінде білімдерді автоматты түрде шығару механизмін жүзеге асыру.

Жұмыстың ғылыми жаңалығы:

1. Қолданыстағы әдістерден (TableMiner++, QTLTableMiner++, MantisTable) артықшылығы бар жетілдірілген итеративті Semantic Table Interpretation (STI) тәсілі ұсынылды, ол:

1.1. субъект-бағандар мен предикаттарды дәлірек анықтауға арналған кеңейтілген эвристикалық ережелерді қолданумен;

1.2. алынған RDF-триплеттер негізінде білім графтарын автоматты түрде толықтыру және жаңарту механизмін енгізумен ерекшеленеді.

2. Халықаралық онтологиялармен (WHOW, HY_Features) сәйкестендірілген және келесі модульдермен толықтырылған Қазақстанның жаңа гидро-онтологиялық моделі әзірленді:

2.1. ҚР Су кодексінің нормативтері негізінде суды жіктеуге арналған Water_Regulations_KZ модулі;

2.2. су объектілерін аймақтардың әлеуметтік-экономикалық көрсеткіштерімен байланыстыратын Socio_Geo_KZ модулі.

Қорғауға шығарылатын ғылыми тұжырымдамалар:

1. Кеңейтілген эвристикаларға негізделген, субъект-бағандар мен предикаттарды анықтауды қамтамасыз ететін және кестелерді RDF

форматына дәлірек түрлендіруге мүмкіндік беретін, F1 метрикасы бойынша TableMiner++, QTLTableMiner++ және MantisTable әдістерінен жоғары нәтижелер көрсететін, білім графтарын автоматты түрде толықтыру және жаңарту механизмі бар итеративті Semantic Table Interpretation (STI) тәсілі.

2. Халықаралық стандарттармен (WHOW, HY_Features) сәйкестендірілген және Water_Regulations_KZ мен Socio_Geo_KZ модульдерімен кеңейтілген, SWRL негізіндегі ереже-бағдарланған логикалық қорытынды шығаруды қолдайтын, «су объектілері ↔ нормативтер ↔ бақылаулар ↔ аймақтар» байланыстарын автоматты түрде анықтауға мүмкіндік беретін Қазақстанның гидро-онтологиялық моделі.

Зерттеу объектісі мен пәні.

– Іле-Балқаш бассейнінің жерүсті сулары туралы гетерогенді дереккөздерде ұсынылған деректер (мониторингтік кестелер, нормативтік құжаттар, статистикалық есептер, геокеңістіктік деректер).

– Гетерогенді дереккөздер негізінде су доменінің білім графын қалыптастыруды және кеңейтуді қамтамасыз ететін кестелік деректерді семантикалық интерпретациялау әдістері мен онтологиялық модельдеу тәсілдері.

Зерттеу әдістері:

– халықаралық стандарттармен сәйкестендіруді қамтитын онтологиялық модельдеу және онтологиялық теңестіру әдістері;

– әзірленген STI итеративті тәсілін қоса алғанда, кестелік деректерді семантикалық интерпретациялау әдістері;

– SWRL негізіндегі білім графтарында ереже-бағдарланған логикалық қорытынды шығару әдістері;

– Semantic Web стандарттары аясында кеңістіктік-уақыттық деректерді формализациялау әдістері.

Зерттеудің теориялық маңыздылығы. STI итеративті үдерісін бағаналарын анықтау, предикат-кандидаттарды және атрибуттық бағандарды айқындау, кейін онтологиялық теңестіру және ережелер негізінде білім шығару кезеңдерінің формализацияланған бірізді жүйесі ретінде теориялық негіздеу.

Зерттеудің практикалық маңыздылығы. PDF/Веб/Excel форматтарындағы кестелерден су ресурстары туралы білім графын автоматты түрде толықтыратын және RDF стандарттары негізінде деректердің интероперабельдігін қамтамасыз ететін «Кесте → RDF → GraphDB → SPARQL → Визуализация» қайта өндірілуі мүмкін конвейерді әзірлеу.

Ғылыми нәтижелердің сенімділігі. Алынған нәтижелердің сенімділігі Semantic Web формализацияланған стандарттарын пайдалануымен, әзірленген онтологиялық модельдің халықаралық онтологиялармен теңестірілуімен және ұсынылған STI тәсілінің қолданыстағы әдістермен эксперименттік салыстырылуы нәтижесінде сапа метрикалары бойынша артықшылық көрсетуімен қамтамасыз етіледі. Нәтижелердің негізділігі Іле-Балқаш бассейнінің нақты деректерінде модельдің апробациясымен, қалыптастырылған білім графының логикалық келісімділігімен және зерттеу

нәтижелерінің рецензияланатын ғылыми басылымдарда жариялануымен дәлелденеді.

Тақырыптың ғылыми-зерттеу бағдарламаларының жоспарларымен байланысы. Диссертациялық жұмыс ҚР ҒЖБМ (қаржыландыру көзі – ҚР ҒЖБМ Ғылым комитеті) ғылыми-зерттеу жобалары аясында орындалды:

– 2021–2023 жылдарға арналған AP09261344 гранттық қаржыландыру жобасы «Геоақпараттық жүйелерді ақпараттық қамтамасыз ету үшін гетерогенді дереккөздерден геокеңістіктік объектілерді автоматты түрде шығару әдістерін әзірлеу»;

– 2024–2026 жылдарға арналған BR24993001 бағдарламалық-нысаналы қаржыландыру жобасы «Қазақ тілін және технологиялық прогресті қолдау үшін үлкен тілдік модельді (LLM) құру».

Жұмыстың апробациясы. Диссертациялық жұмыстың негізгі нәтижелері халықаралық және шетелдік ғылыми конференцияларда, ғылыми семинарларда баяндалып, талқыланды:

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

– Әл-Фараби атындағы ҚазҰУ Ақпараттық технологиялар және жасанды интеллект факультеті Информатика кафедрасының ғылыми семинарында.

Диссертациялық жұмыстың құрылымы мен қысқаша мазмұны. Диссертация кіріспеден, төрт бөлімнен, қорытындыдан және пайдаланылған әдебиеттер тізімінен тұрады.

Кіріспеде тақырыптың өзектілігі негізделіп, зерттеудің мақсаты мен міндеттері тұжырымдалған, зерттеу нысаны, пәні және әдістері айқындалған, жұмыстың ғылыми жаңалығы, теориялық және практикалық маңыздылығы ашып көрсетілген.

Бірінші бөлімде су ресурстарын мониторингтеуде онтологиялық тәсілдің теориялық негіздері қарастырылып, гидрология саласындағы қолданыстағы онтологияларға талдау жасалған, сондай-ақ гидрологиялық модельдердегі семантикалық қорытындылау мен жаңа білімді шығару әдістері сипатталған.

Екінші бөлімде кестелік деректерді семантикалық интерпретациялау тәсілдері, STI-жүйелерінің тиімділігін бағалау әдістері зерттеліп, Қазақстанның су домені бойынша гетерогенді дереккөздерге жинақтау және талдау жүргізілген, сонымен қатар құрылымдалмаған көздерден кестелерді шығару құралдары қарастырылған.

Үшінші бөлімде онтологиялық-бағытталған геоақпараттық модельді қалыптастыру әдіснамасы әзірленіп, оның көпдеңгейлі және модульдік архитектурасы ұсынылған, бақылау, гидрографиялық объектілер және ұлттық нормативтер модульдері сипатталған, итеративті STI әдісі мен SWRL негізіндегі ереже-бағытталған қорытындылау механизмі ұсынылған.

Төртінші бөлімде әзірленген тәсілдің эксперименттік валидация нәтижелері келтіріліп, кестелерді семантикалық интерпретациялаудың

қолданыстағы әдістерімен салыстырмалы талдау жасалған, SWRL-ережелерін қолдану нәтижесінде алынған жаңа аксиомалар көрсетіліп, GraphDB ортасында интеграцияланған гидролого-әлеуметтік онтологияның іске асырылуы демонстрацияланған.

Қорытындыда диссертациялық зерттеудің негізгі нәтижелері мен тұжырымдары берілген.

Зерттеу нәтижелері бойынша жарияланымдар. Диссертациялық зерттеу нәтижелері 6 ғылыми еңбекте жарияланған. Оның ішінде: 2 мақала Web of Science және Scopus деректер базасына енгізілген халықаралық ғылыми журналдарда, 1 жұмыс халықаралық және республикалық конференциялар материалдарында, 3 мақала ҚР ҒЖБМ Білім және ғылым саласын бақылау комитеті ұсынған журналдарда, және 1 авторлық куәлік диссертациялық зерттеу нәтижелерін практикалық енгізу бойынша алынған.:

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. Авторлық куәлік: Оспан Әсел, Мансурова Мадина Есимхановна. Гетерогенді дереккөздерден алынған кестелерді семантикалық талдау арқылы білім графтарын толықтыруға арналған итеративті алгоритм. ЭЕМ-ге арналған бағдарламаны мемлекеттік тіркеу туралы куәлік № 39296, 2023 жылғы «28» қыркүйек.