Brief information about the project

Name of the project	AP09261344 «Development of methods for automatic extraction of geospatial objects from heterogeneous sources for information support of geographic information systems» (0121PK00388)
Relevance	The relevance of the work lies in the inevitable need for effective processing and integration of geospatial data from various open sources, especially from text web tables. Complexities associated with incompatible formats and lack of semantics hinder the integration process and can lead to the loss of key decision-making information. The development of methods for automatically extracting geospatial features and their attributes from web text tables represents an important step toward simplifying this process, allowing for more efficient use of geographic information systems and improved data analysis for informed decision making in a geographic context.
Purpose	The goal of the project is to develop methods for automatically extracting geospatial objects and associated non-spatial attributes from heterogeneous open data sources, namely web text tables.
Objectives	To achieve the project goal, the following tasks need to be solved: 1. Research and development of semantic methods for extracting and interpreting geospatial objects and their quantitative and qualitative descriptions from text web tables as sets of attribute-value pairs. 2. Research and development of methods for integrating and aligning extracted geospatial data based on open geoinformation resources of the Semantic Web. 3. Consolidation of the created methods and algorithms into a single technology, the basis of which is a schemaless distributed NoSQL model. **4.** Prototyping a software product based on the developed technology. Creation of web services for parsing and extracting geospatial information from websites in the "Tourism" and "Emergencies" domains.
Expected and achieved results	According to the results of the project: − at least 3 (three) articles and (or) reviews will be published in peer-reviewed scientific publications indexed in the Science Citation Index Expanded of the Web of Science database and (or) having a CiteScore percentile in the Scopus database of at least 50 (fifty); − and 2 articles in the proceedings of international conferences indexed in the Scopus database, for example, Computational Collective Intelligence Conference; − at least 3 (three) articles or reviews in a peer-reviewed foreign or domestic publication recommended by KOKSON RK;

	− and 1 monograph in the Kazakh publishing house (Kazakh University); − you will receive a copyright certificate of state registration of rights to the copyrighted object. As a result of the completion of the project, testing of software technologies for the successful use of technology for automatic extraction of geospatial objects, it is planned to develop scientific and technical documentation. Results achieved: - Intelligent methods for extracting data from text tables as sets of attribute-value pairs, methods for analyzing the physical, functional and logical structure of web tables and corresponding parsers for recognizing web tables depending on the type of input data have been developed. - Methods have been developed for semantic interpretation of geodata, including distributed loading of data into an unstructured key-value storage, semantic transformation of data into an object representation based on an ontological approach, determination, and clarification of the coordinate reference of extracted geodata using extracted data. - A technology has been developed for automatically extracting geoinformation from text tables on the Web, a cloud-based distributed infrastructure with the consolidation of the created methods and algorithms into a single service.
Research team members with their identifiers (Scopus Author ID, Researcher ID, ORCID, if available) and links to relevant profiles	1. Mansurova Madina Yesimkhanovna - Candidate of Physical and Mathematical Sciences, Associate Professor, Head of the Department of Artificial Intelligence and Big Data of KazNU. al-Farabi, leading researcher at Kazakh National University. al-Farabi. Scopus H-index =5, Web of Science H-index = 2, publications indexed in Scopus – 64, total number of citations – 91. 2. Nugumanova Aliya Bagdatovna – PhD, Director of the Big Data and Blockchain Technologies Research Center Astana IT University. Scopus Author ID: 55864815200, Orcid ID: 0000-0001-5522-4421, h-index=5. 3. Vladimir Borisovich Barakhnin – higher education, graduated from Novosibirsk State University, ResearchGate: A-5856-2014, ORCID: https://orcid.org/0000-0003-3299-0507, SCOPUS: 6508258628. 4. Shomanov Adai Sakenovich – PhD, employee of Nazarbayev University, Scopus Author ID: 57195543732, h-index Scopus = 4. 5. Ospan Asel Galymzhankyzy – master's student, senior lecturer at the Faculty of Information Technologies. Scopus

	Author ID: 57238489800, ORCID ID: 0000-0002-1860-6997, h-index=1.
List of publications with links to them	1. 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. 2. Kadyrbek N, Mansurova M, Shomanov A, Makharova G. The Development of a Kazakh Speech Recognition Model Using a Convolutional Neural Network with Fixed Character Level Filters. Big Data Cogn Comput. 2023;7(3):132. doi:10.3390/bdcc7030132. 3. 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. 4.K. Bauyrzhan, M. Madina and O. Assel, "Fine-Tuning the Wav2vec2 Model for Kazakh Speech: A Study on a Limited Corpus," *2023 IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, 2023, pp. 124-128, doi: 10.1109/SIST58284.2023.10223504. 5.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. 6.Mansurova M, Ospan A, Kakimzhanov Y, Resnik B, Tyulyubayev D. Development of an Application for Monitoring and Analyzing the Dynamics of the Tuyuk Su Mountain Glacier. SIST 2022 International Conference on Smart Information Systems and Technologies. https://sist.astanait.edu.kz/wp-content/uploads/2022/05/conference-programme-129.pdf. Published 2022. 7.Mansurova M, Barakhnin V, Kyrgyzbayeva M, Kadyrbek N. Named Entity Extraction Model Based on the Random Walk Method. In: 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST). IEEE; 2021. https://ieeexplore.ieee.org/document/9465992. 8.Ospan A, Mansurova M, Kakimzhanov E, Aldakulov B. KazRivDyn: Toolkit for Measuring the Dynamics of Kazakhstan Rivers with Graphics Based on Google Earth Engine. In: 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST). IEEE; 2021. https://ieeexplore.ieee.org/document/9465902. 9.Akhmed-Zaki D, Mansurova M, Yertuyak A, Chikibayeva D. Development of Web Application for Visualizing City Emergencies. 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST). IEEE; 2021. doi:10.1109/SIST50301.2021.9465919.

	10.Meiran Zhiyenbayev, Assel Ospan, Nadezhda Kunicina, Madina Mansurova, Roman Titkov. Systematic data procurement in an owl-embedded information and analytical framework for the monitoring of water resources in the Ile-Balkhash basin. Scientific Journal of Astana IT University, ISSN (P): 2707-9031 ISSN (E): 2707-904X, Volume 15, September 2023. 11.Zhiyenbayev M, Ospan A, Mansurova M. ETL Process for Water Resources and Demographics Data: An Open Source Data Processing Tools and Visualizations. Vestn Nats Inzh Akad Respub Kaz. 2023;(88):38-48. doi:10.47533/2023.1606-146X.4. 12.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. 13. Ospan A, Mansurov 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. **Monograph:** 2. M. E. Mansurova. Advanced models and methods of Text Mining: monograph. - Almaty: Kazakh University, 2023. - 112 . ISBN 978-601-04-6499-5
Patents	Copyright certificates: 1. Ospan Asel, Mansurova Madina Yessimkhanovna.An iterative algorithm for semantic analysis of tables from heterogeneous sources to replenish knowledge graphs. Certificate of state registration of the computer program No. 39296 dated September 28, 2023. 2. Mansurova Madina Yessimkhanovna, Kadyrbek Nurgali, Dosanov Bekzhan, Kyrgyzbayeva Marzhan, Tyulepberdinova Gulnur.The pipeline of preprocessing texts in the Kazakh language. No. 17792 dated May 21, 2021. 3. Mansurova Madina Yessimkhanovna, Chikibaeva Darya Yuryevna, Tyulepberdinova Gulnur Alpykyzy. An algorithm for extracting named entities from news sources in the Kazakh language based on bi-LSTM. No. 17402 dated May 12, 2021.
Video: https://youtu.be/CF0ie1zDX1E	