УДК 20.53.19
https://doi.org/10.47533/2023.1606-146X.4

*M. M. ZHIYENBAYEV, A. G. OSPAN, M. E. MANSUROVA**

Department of Artificial Intelligence and Big Data, Faculty of Information Technology,
Al-Farabi Kazakh National University
Almaty, Kazakhstan
meiran1991@gmail.com; asselyaospan@gmail.com; mansurova.madina@gmail.com

ETL PROCESS FOR WATER RESOURCES AND DEMOGRAPHICS DATA: AN OPEN SOURCE DATA PROCESSING TOOLS AND VISUALIZATIONS

Open data portals have become increasingly popular in recent years as more and more governments, organizations, and businesses recognize the benefits of making data publicly available for researchers, analysts and decision-making platforms. Many governments have launched their portal at the national, regional, and local levels. In addition to traditional government datasets such as demographics, finance, and transportation, open data portals are increasingly providing access to a broader range of data types, including environmental data, health data, and social media data. Many open data portals are offering improved tools for data analysis and visualization, making it easier for users to extract insights from the data. Increased collaboration and engagement between government, businesses, and the public, with features such as forums and hackathons. The purpose of this research is to find a way to automate data collecting and visualization processes by using different open source technologies and promoting Open Data Portal by using environmental data, especially water resources and pollution data from the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan, Bureau of National Statistics.

Key words: *data package, data portal, water pollution index, open data, csv.*

Introduction. Open data portals are online platforms that provide access to government data in a structured, standardized, and machine-readable format. Here are some of the key reasons why open data portals are important for countries:

− Transparency and Accountability: Open data portals increase transparency and accountability by making government data available to the public. Citizens can use this information to hold government officials accountable for their actions and decisions [1].

− Economic Development: Open data portals can stimulate economic development by providing businesses with access to valuable data that can inform their decisions. This can lead to the creation of new businesses and job opportunities [2].

− Improved Service Delivery: Open data portals can improve service delivery by making government data available to service providers and other stakeholders. This can help service providers to better understand the needs of their clients and to design more effective programs and services [1].

− Innovation: Open data portals can stimulate innovation by providing entrepreneurs, researchers, and developers with access to data that can be used to create new products and services. This can lead to new business opportunities, increased competitiveness, and economic growth [3].

− Better Decision Making: Open data portals can enable better decision making by providing policymakers with access to data and analytics. This can help them to design more effective policies and programs that meet the needs of citizens [1].

* E-mail корреспондирующего автора: mansurova.madina@gmail.com

Overall, open data portals are important tools for countries that want to promote transparency, accountability, economic development, innovation, and better decision making. By making government data available to the public in a structured and accessible format, open data portals can help to drive positive change and improve the lives of citizens. Figure 1 depicts the open data portals by Open Knowledge Foundation https://okfn.org/, there are 597 open data portals from around the world. Unfortunately, in Kazakhstan, we have only few portals compared to other countries. More open portals in the Republic of Kazakhstan will allow us to make more informed data-driven decisions.

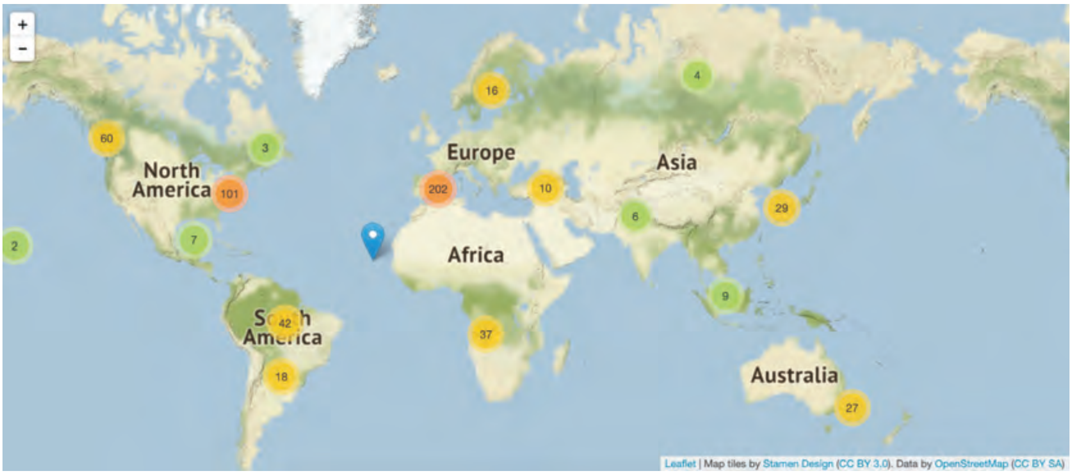

+
-
60
3
North America
101
2
7
16
Europe
202
10
Africa
4
Asia
6
29
9
42
South America
18
37
Australia
27
Leaflet | Map tiles by Stamen Design (CC BY 3.0). Data by OpenStreetMap (CC BY SA)

Figure 1 **–** Open Data Portals by Open Knowledge Foundation (dataportals.org [4])

For the water resources of Kazakhstan, there are also open data portals, which collect socio-economic indicators, data on Kazakhstan's water resources, air pollution and surface water pollution, and data on industries. But despite the fact that all data is in the public domain, all information is stored in unstructured or semi-structured sources, such as text documents, excel files or PDF documents. Storing data in such formats makes it difficult to understand the situation and complicates readability. To do this, this paper presents dashboards for visualizing data on the water resources of Kazakhstan and their relationship with socio-economic indicators. Since water resources are an important part of the country's economy and agro-industry, 8 main water intake basins with hydro-energy resources, as well as rivers and water pollution indices, were selected for visualization. (Figure 2). Architecture and electronic graphics are built on the basis of these data.

ETL Design and Architecture. This paper takes you through the entire ETL process, as shown in Figure 3, data management and visualization. ETL is the process of extracting data from one or more sources, transforming it to fit the target schema, and loading it into a target data warehouse. ETL is a complex and time-consuming process that requires careful planning and execution [5]. First part of this process we explain the tools and packages to make data powerful and readable, in addition, we provide you schema for processing data from the source to the structured form with publicly available tools. When our data is clean

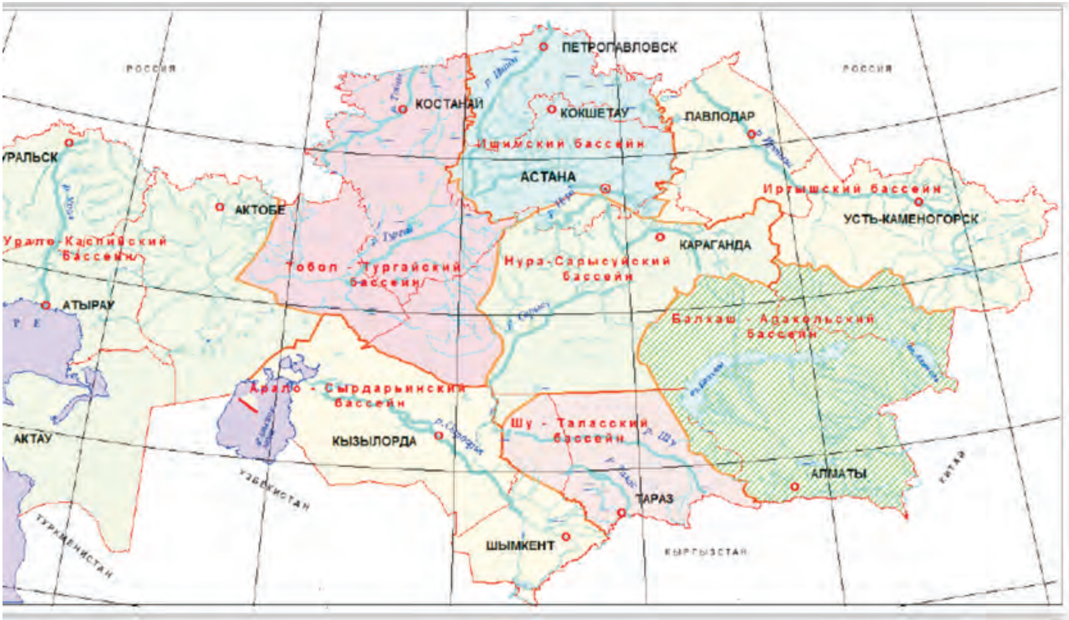


Figure 2 – 8 water basins of Kazakhstan: 1. Aral-Syrdarya, 2. Ile-Balkhash, 3. Irtysh, 4 Ishim, 5. Nura-Sarysu, 6. Tobol-Turgay, 7. Ural-Caspian, 8. Shu-Talas

and structured in commonly used data format, as an example we came up with comma separated values(CSV) files. There are huge advantages using CSV file format:

– easy to read and write - CSV is an easy-to-read and - write file format for data exchange between systems and applications [6];

– lightweight - CSV files are lightweight, and their small file size makes them easy to share and transfer over the internet [7];

– compatible with a wide range of applications - CSV is a universal format that can be easily imported and exported by a wide range of applications, including databases, programming languages, and spreadsheet software [8];

– preserves data structure - CSV format preserves the data structure, including column headers and data types, making it easy for users to understand and work with the data [9].

– open standard - CSV is an open standard format that is widely used for data exchange because it is simple, widely supported, and platform-independent [10].

ETL process for generation structured datasets is built using Python programming language with automation of some open source technologies. During the cleaning process, the major role is played by the dataflows [11] library which has all necessary methods to clean, parse, pivot data. The source data comes from the Agency for Strategic planning and reforms of the Republic of Kazakhstan, Bureau of National Statistics https://stat.gov.kz/. Analyzing the documents, they are not standardized and structured, mainly in excel format with different structure and data formats which are not structured to use in building dashboards and doing some analytics. So our goal is to build a data pipeline which has collecting, cleaning, formatting and packaging steps which are fully automated and scheduled depending on the source published dates.

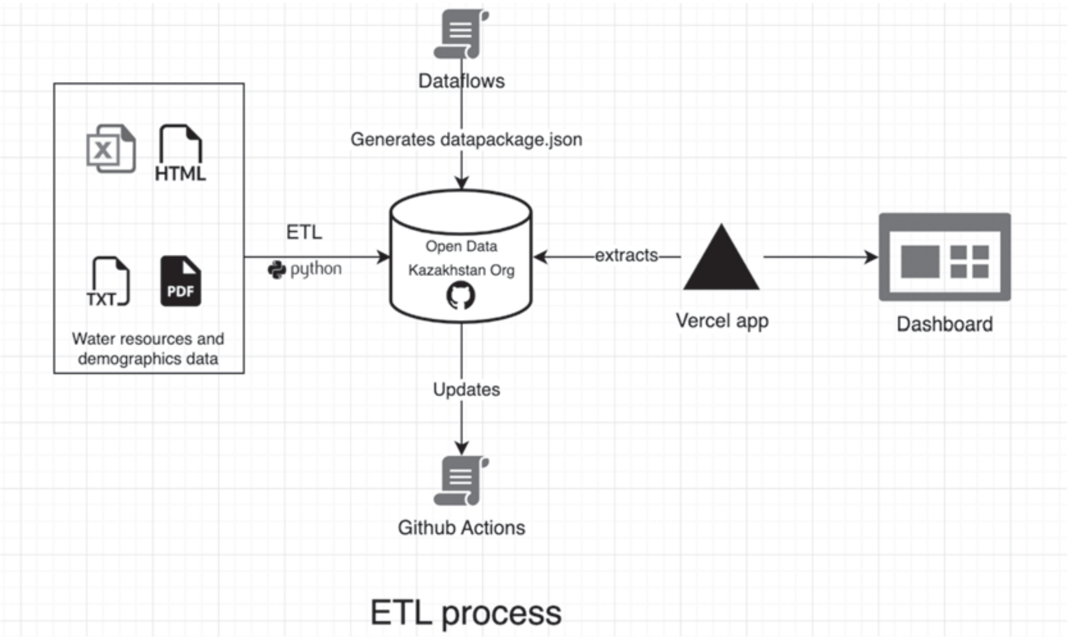
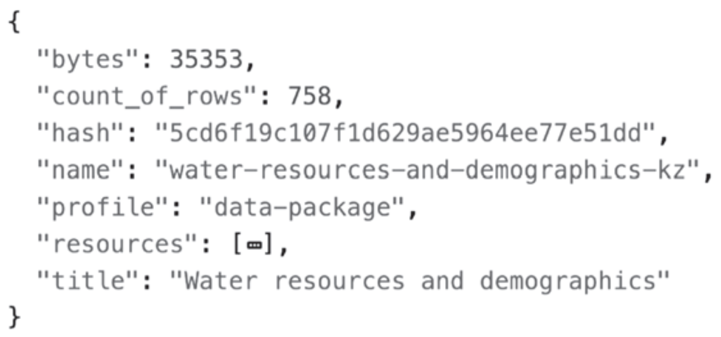


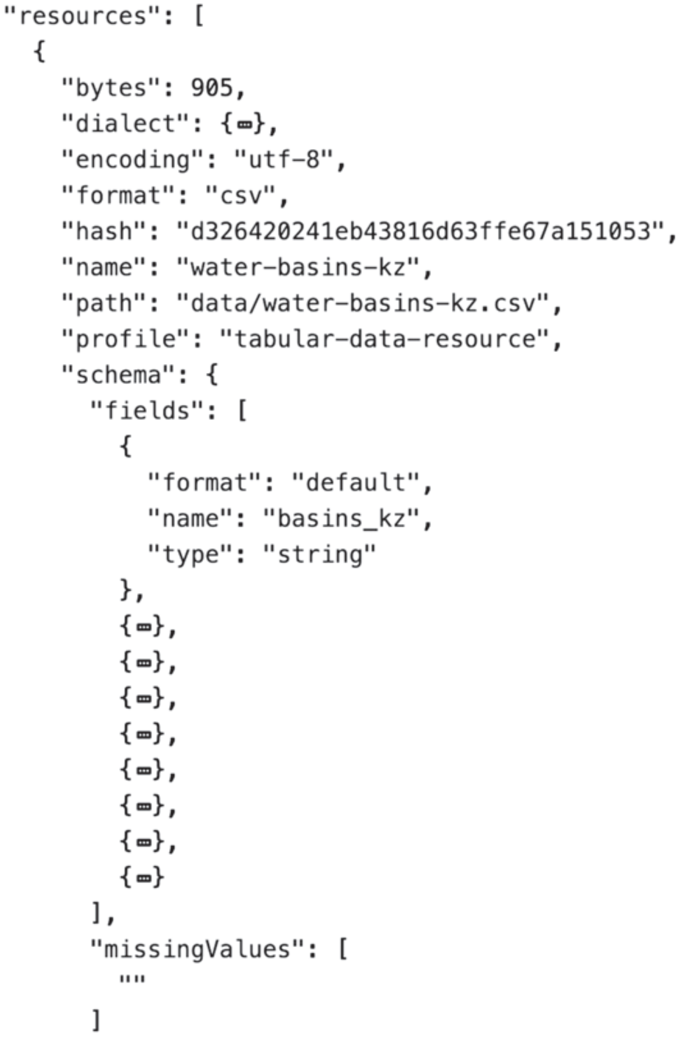
Figure 3 – ETL process for water resources and demographics data

When building our dataset, we use lightweight yet comprehensive data standards as Data Package [12]. Basic container format for describing a collection of data in a single “package”. It provides a basis for convenient delivery, installation and management of datasets. A Data Package can contain any kind of data. At the same time, Data Packages can be specialized and enriched for specific types of data so there are, for example, Tabular Data Packages for tabular data, Geo Data Packages for geo data etc. Figure 4 shows metadata json file, you can find all necessary information about the data such as author, license, published date, total of rows, profile, hash and metadata about each resource with format, encoding, name, title schema describing each field, as shown in Figure 5. The Data Package is supported also by Open Knowledge Foundation and its specification is updated and posted in channels, posts [13].



```
{
  "bytes": 35353,
  "count_of_rows": 758,
  "hash": "5cd6f19c107f1d629ae5964ee77e51dd",
  "name": "water-resources-and-demographics-kz",
  "profile": "data-package",
  "resources": [⋯],
  "title": "Water resources and demographics"
}
```

Figure 4 – This is the sample generated metadata about the package.



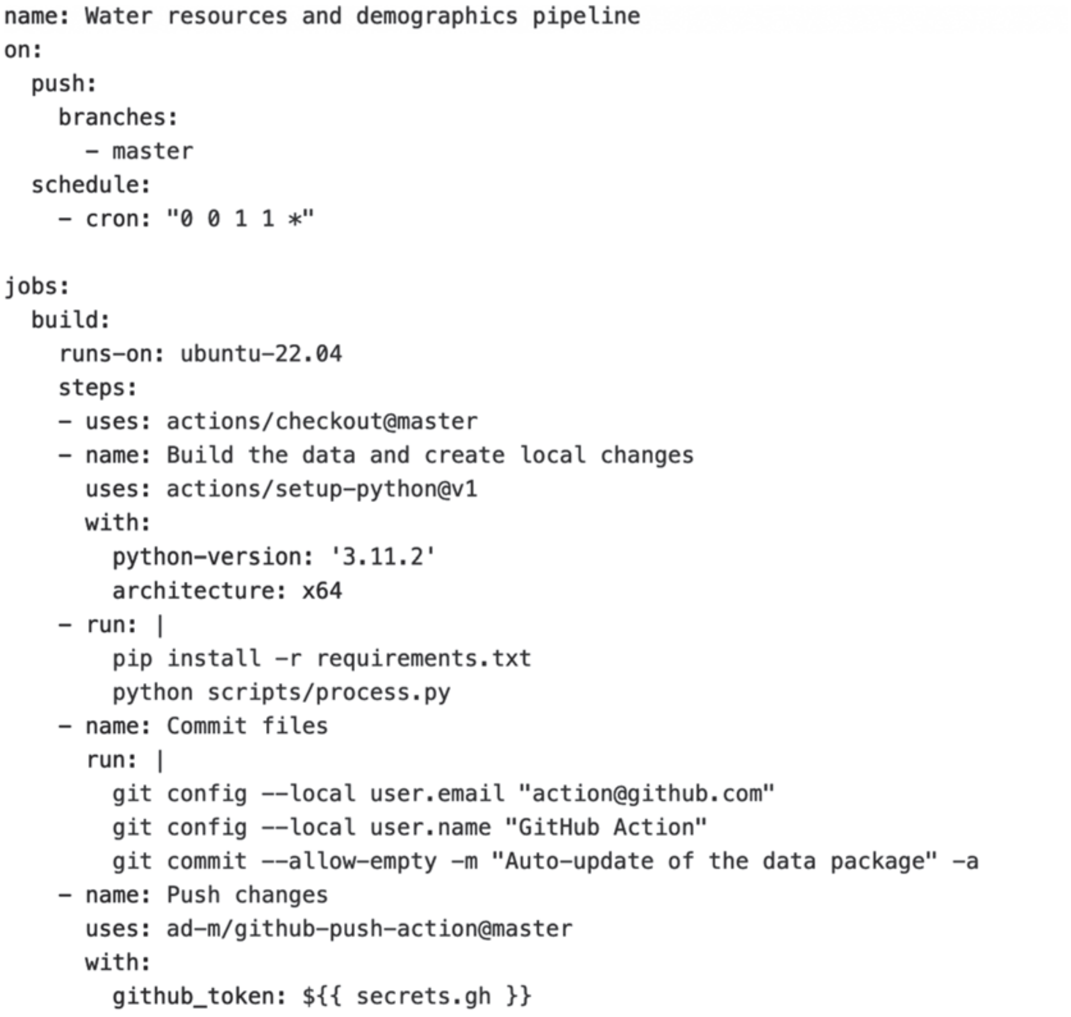
```
"resources": [
  {
    "bytes": 905,
    "dialect": {⋯},
    "encoding": "utf-8",
    "format": "csv",
    "hash": "d326420241eb43816d63ffe67a151053",
    "name": "water-basins-kz",
    "path": "data/water-basins-kz.csv",
    "profile": "tabular-data-resource",
    "schema": {
      "fields": [
        {
          "format": "default",
          "name": "basins_kz",
          "type": "string"
        },
        {⋯},
        {⋯},
        {⋯},
        {⋯},
        {⋯},
        {⋯},
        {⋯},
        {⋯}
      ],
      "missingValues": [
        ""
      ]
```

Figure 5 – This is the sample generated metadata about the resources (github.com/open-data-kazakhstan/water-resources-and-demographics[14])

Data storage and management. Second part of the process is to keep the data up to date and publicly available. In order to accomplish this step, we are using and enhancing Open Data Kazakhstan organization [15] in the GitHub platform which collects, cleans, structures and keeps data from different formats. Currently, this organization has several core datasets such as “Decent work indicators”, “Gross domestic product per capita using production method by regions”, “Average per capita nominal cash income of the population”, “Demographics in Kazakhstan by region, age and sex.”, “Covid statistics” and so on. In addition, this organization develops a “School of Data” portal where students and enthusiasts about the data can join and solve real world issues by helping our country to prepare Data Engineers. For these reasons we came up with the solution to keep data in Open Data Kazakhstan organization in Github repository since all our sources are public and open

[16]. For this paper, we are using the “Water resources and Demographics” repository under the organization and make it public so everyone can access it and get some insight from our dashboards and use it for their papers and future works. Lastly, the reason behind publishing datasets into Open Data Kazakhstan is to promote open data portals in our country, to help researchers, analysts, and students to do some analysis with structured and clean datasets.

In order to keep data up to date we added Github Actions [17] which runs periodically depending on the data, the configuration file as shown in Figure 6. GitHub Actions is a powerful tool that allows developers to automate their workflows and streamline their development processes. Here are some advantages of using GitHub Actions for updating data such as automated data updates, version control, collaboration, integration with other tools and open-source community [18].



```
name: Water resources and demographics pipeline
on:
  push:
    branches:
      - master
  schedule:
    - cron: "0 0 1 1 *"

jobs:
  build:
    runs-on: ubuntu-22.04
    steps:
    - uses: actions/checkout@master
    - name: Build the data and create local changes
      uses: actions/setup-python@v1
      with:
        python-version: '3.11.2'
        architecture: x64
    - run: |
        pip install -r requirements.txt
        python scripts/process.py
    - name: Commit files
      run: |
        git config --local user.email "action@github.com"
        git config --local user.name "GitHub Action"
        git commit --allow-empty -m "Auto-update of the data package" -a
    - name: Push changes
      uses: ad-m/github-push-action@master
      with:
        github_token: ${{ secrets.gh }}
```

Figure 6 – This is the generated actions.yml for Github actions (workflows/actions.yml [15])

Data visualization. In addition, we build some graphs out of data, so it can be useful to see some trends in water consumption and resources. Dashboard [19] is built by using the Vercel [20] platform which provides the speed and reliability during the inspiration of our researchers to quickly build frontend applications. Vercel is a platform that provides a serverless development environment for building web applications, including data visualizations. Here are some benefits of using Vercel for building visualizations such as high performance by loading visualizations quickly,since it is based on serverless architecture, your visualizations can scale automatically to handle large amounts of data, a simple deployment process that allows you to easily deploy your visualizations to the web, also supports collaboration features which allow to work with other developers on your visualizations, making it easy to collaborate on projects and share code. The codebase is also available publicly on Github [21]. In addition, the reason we are building visualization on Vercel, it integrates with popular data visualization libraries such as D3.js and Plotly [22]. It is publicly available on the web so researchers, analysts, data enthusiasts can take a look at the water resources and consumptions by basins, rivers and water pollution indicators.

Dashboard and processed dataset are synchronized together so it will be updated automatically when changes will be applied to the dataset repository, so it will be updated. During the implementation of this project, we proposed only open source tools and libraries, so anyone can create data pipelines and dashboards freely by enriching our publicly open organization. Additionally, there are some graphs on average annual water consumption, water and energy resources. These are some samples taken from the dashboard such as water resources in cure meters by basins, as shown in Figure 7, share of river (Figure 8) and water pollution index (Figure 9).

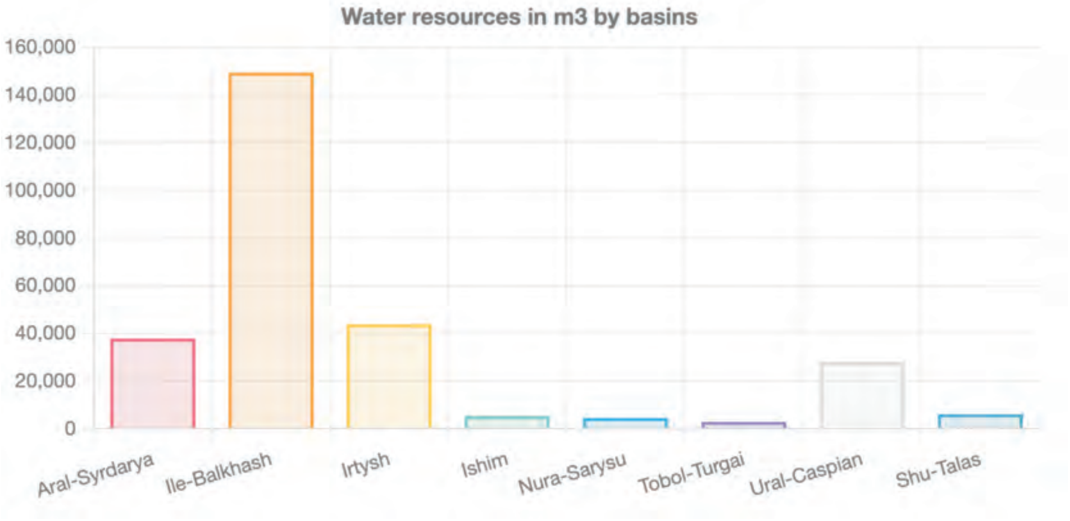


Figure 7 **–** Water resources in m3 by basins - Republic of Kazakhstan (https://water-resources-and-demographics-dashboard.vercel.app/[19])

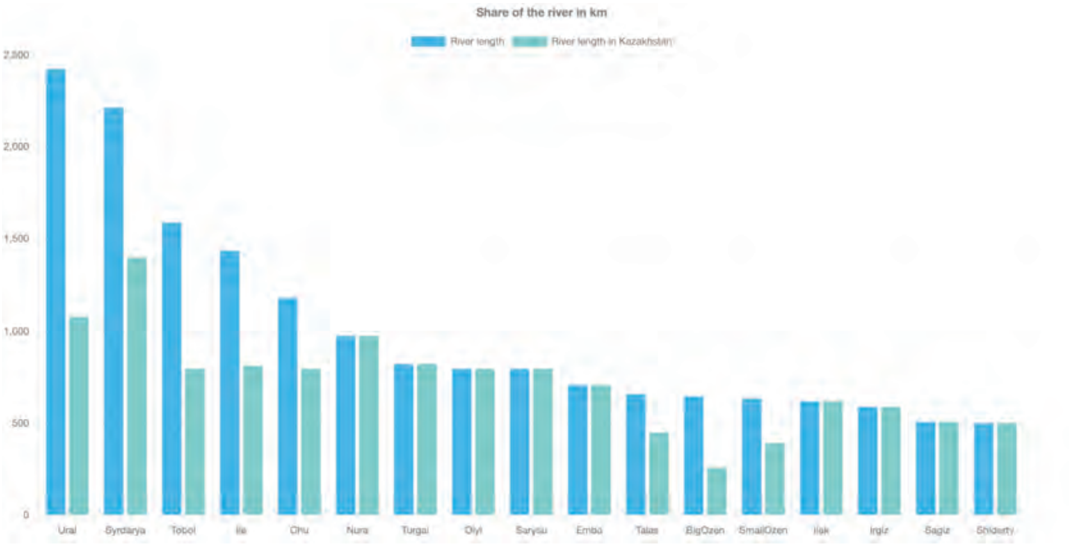


Figure 8 – Share of the river in kilometers - Republic of Kazakhstan (https://water-resources-and-demographics-dashboard.vercel.app/[19])

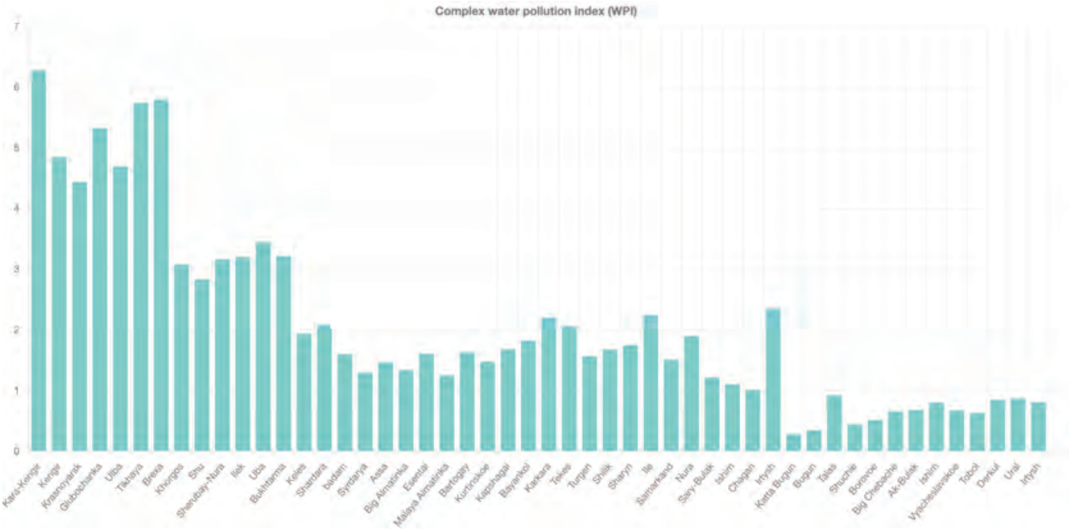


Figure 9 – Complex water pollution index(WPI) - Republic of Kazakhstan (https://water-resources-and-demographics-dashboard.vercel.app/[19])

Results. To visualize data on water intake basins, rivers and the water pollution index, the following steps were performed:

1. Collection of data on socio-economic indicators from open sources, such as the Bureau of National Statistics of the Republic of Kazakhstan (stat.gov), 8 water intake basins and rivers, surface water pollution index (kazhydromet.kz);

2. Data conversion from unstructured and semi-structured data to a structured format (CSV);

3. Placement of data on Open Data Kazakhstan on the GitHub platform, which collects, cleans, structures and stores data from different formats;

4. Graphing for better understanding and data analysis.

Further, based on the constructed graphs, the user can do a comparative analysis. In our research area, we see that the largest water intake basin is the Ile Balkhash basin, which is located in Almaty, Zhambyl and Karaganda regions. Also, these regions are the most densely populated regions, also the Karaganda region is the most industrial region. All this leads to the fact that it is from the Ile-Balkhash basin that there will be the most water intake for the agricultural and household needs of the population, for industries and agriculture.

On the basis of the 3rd graph (Fig. 8) it is possible to trace the pollution of water bodies. Here we can conclude that the largest WPI of water falls on those water bodies that are closer to industrial plants, and the growth of the population of the region also has a direct impact on the WPI.

Based on the results of this work, we can conclude that by building dashboards for data on water resources and socio-economic indicators, it is possible to better analyze the data and identify correlations between these indicators. Further on this work, it is planned to add data on the level and flow of water, on the incidence of the population living close to the water intake populations, on industries using hydro-energy resources and changes in WPI over time intervals. All this in the future will show us a clearer picture of the impact of water resources in general on the overall social and economic indicators of the country. All the datasets which are used for the analysis are published on “Open Data Kazakhstan” open data portal [23] which provides enhanced open data systems for public use.

Conclusion. We demonstrated the power of open data and tools where we can create a data pipeline and dashboards using open source projects. Contributing to open data portal would help other researchers to make a quality data analysis and decision making based on structured and clean data. As we discussed in this paper, increasing data portals at the national, regional and local levels could help the public to get more insights from the data. In addition, we are not only providing access to environmental data which plays a key role nowadays when we have a water shortage. This paper would trigger other researchers to collect and do analysis on other parts of data such as media, health, transportation.

Acknowledgement. This work was funded by Committee of Science of Republic of Kazakhstan AR09261344 “Development of methods for automatic extraction of geospatial objects from heterogeneous sources for information support of geographic information systems” (2021-2023).

REFERENCES

1 Open Data Handbook - Transparency and Accountability - https://opendatahandbook.org/guide/en/why-open-data/#transparency-and-accountability

2 World Bank - Open Data for Economic Growth - https://www.worldbank.org/en/topic

3 European Data Portal - Open Data as a Driver of Innovation - https://data.europa.eu/en.

4 A Comprehensive List of Open Data Portals from Around the World by Open Knowledge Foundation - https://dataportals.org/

5 Kimball, R. & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition.

6 E. Rössel, M. Matten, & M. Westenberger (2018). Software Tools for Academic Research: A Guidebook for Social Scientists. Springer International Publishing.

7 J. P. Grimes & D. D. Jorgensen (2016). Introductory Statistics: A Conceptual Approach Using R. CRC Press.

8 D. Duggan & M. Gruen (2015). Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph. Apress.

9 P. Baumann & K. Janowicz (2015). GeoSemantic Technologies for Intelligent Transportation Systems. Springer International Publishing.

10 S. L. Morgan & C. J. Winship (2015). Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press.

11 Dataflows library - https://github.com/datahq/dataflows

12 Data package standart - https://frictionlessdata.io/projects/#individual-file

13 Zhiyenbayev M. (2017). Data Package v1 Specifications. What has Changed and how to Upgrade. Open Knowledge Lab. https://okfnlabs.org/blog/2017/10/11/upgrade-to-data-package-specs-v1.html

14 Water Resources and Demographics Republic of Kazakhstan - https://github.com/open-data-kazakhstan/water-resources-and-demographics

15 Github repository for dataset - https://github.com/open-data-kazakhstan/water-resources-and-demographics

16 Open data - https://en.wikipedia.org/wiki/Open_data

17 Github actions - https://docs.github.com/en/actions

18 GitHub Actions Documentation - https://docs.github.com/en/actions

19 Water resources and demographics dashboard - https://water-resources-and-demographics-dashboard.vercel.app/

20 Vercel platform - https://vercel.com/

21 Vercel app for dashboard - https://github.com/Mikanebu/water-resources-and-demographics-dashboard

22 Supported Frameworks on Vercel - https://vercel.com/docs/frameworks

23 Open Data Kz - https://www.opendatagov.kz/

М. М. ЖИЕНБАЕВ, Ә. Ғ. ОСПАН, М. Е. МАНСУРОВА

Жасанды интеллект және Big Data, Ақпараттық технологиялар факультеті,
әл-Фараби атындағы Қазақ ұлттық университеті
Алматы қ., Казақстан

СУ РЕСУРСТАРЫНА ЖӘНЕ ДЕМОГРАФИЯЛЫҚ ДЕРЕКТЕРЛЕРГЕ ӨҢДЕУГЕ АРНАЛҒАН ETL ПРОЦЕСІ: ДЕРЕКТЕРДІ ӨҢДЕУДЕГІ OPEN SOURCE ҚҰРАЛДАР МЕН ВИЗУАЛИЗАЦИЯЛАР

Ашық деректер порталдары соңғы жылдары көбірек танымал болды, өйткені көбірек үкіметтер, ұйымдар және бизнес зерттеушілер, талдаушылар және шешім қабылдау платформалары үшін жалпыға қолжетімді деректердің қолжетімділігінің артықшылықтарын мойындайды. Көптеген үкіметтер ұлттық, аймақтық және жергілікті деңгейлерде өз порталдарын іске қосты. Демография, қаржы және көлік сияқты дәстүрлі мемлекеттік деректер жиындарына қоса, ашық деректер порталдары қоршаған орта, денсаулық және әлеуметтік *медиа деректерін қоса алғанда, деректер түрлерінің кең ауқымына қол жеткізуді қамтамасыз етуде. Көптеген ашық деректер порталдары пайдаланушыларға деректерден түсінік алуды жеңілдететін жақсартылған деректерді талдау және визуализация құралдарын ұсынады. Форумдар мен хакатондар сияқты мүмкіндіктер арқылы үкімет, бизнес және жұртшылық арасындағы ынтымақтастық пен өзара әрекеттесуді арттырыңыз. Бұл зерттеудің мақсаты әртүрлі ашық бастапқы технологияларды пайдалана отырып, деректерді жинау және визуализациялау процестерін автоматтандыру жолын табу және қоршаған орта деректерін, әсіресе Республиканың Стратегиялық жоспарлау және реформалау агенттігінің су және ластану деректерін пайдалана отырып, Қазақстанның, Ұлттық статистика бюросының ашық деректер порталын ілгерілету.*

***Түйін сөздер**: деректер пакеті, деректер порталы, судың ластану индексі, ашық деректер, csv.*

М. М. ЖИЕНБАЕВ, А. Г. ОСПАН, М. Е. МАНСУРОВА

Искусственный интеллект и Big Data, факультет информационных технологий,
Казахский национальный университет имени аль-Фараби, г. Алматы

ETL ПРОЦЕСС ДЛЯ ВОДНЫХ РЕСУРСОВ И ДЕМОГРАФИЧЕСКИХ ДАННЫХ: ИНСТРУМЕНТЫ ОБРАБОТКИ ДАННЫХ С ОТКРЫТЫМ ИСТОЧНИКОМ И ВИЗУАЛИЗАЦИЯ

Порталы открытых данных становятся все более популярными в последние годы, поскольку все больше и больше правительств, организаций и предприятий признают преимущества общедоступности данных для исследователей, аналитиков и платформ для принятия решений. Многие правительства запустили свои порталы на национальном, региональном и местном уровнях. В дополнение к традиционным государственным наборам данных, таким как демография, финансы и транспорт, порталы открытых данных все чаще предоставляют доступ к более широкому диапазону типов данных, включая данные об окружающей среде, данные о здоровье и данные социальных сетей. Многие порталы открытых данных предлагают улучшенные инструменты для анализа и визуализации данных, упрощающие пользователям извлечение информации из данных. Расширение сотрудничества и взаимодействия между правительством, предприятиями и общественностью происходит благодаря таким функциям, как форумы и хакатоны. Целью данного исследования является поиск способа автоматизации процессов сбора и визуализации данных с использованием различных технологий с открытым исходным кодом и продвижения Портала открытых данных с использованием данных об окружающей среде, особенно данных о водных ресурсах и загрязнений от Агентства стратегического планирования и реформ Республики Казахстан, Бюро национальной статистики.

***Ключевые слова**: пакет данных, портал данных, индекс загрязнения воды, открытые данные, csv.*