BABELSTREET

CASE STUDY

Using AI-powered Name Matching for a Central Asian Due Diligence Database

Latvia-based startup ClearPic provides a risk-management platform leveraging a master due diligence business database it has created for Central/Western Asia and the Caspian region. ClearPic is enabling organizations worldwide to comply with anti-money laundering (AML) and sanctions screening regulations. Major international business data providers do not fully cover this region, which prevents European and Western countries from doing business and investing there because they do not have the data to perform sanctions screening and due diligence of potential partners. These companies are obligated to screen all partners to stay in compliance with the Foreign Corrupt Practices Act (FCPA) that prohibit bribing foreign government officials to obtain or retain business.[1]

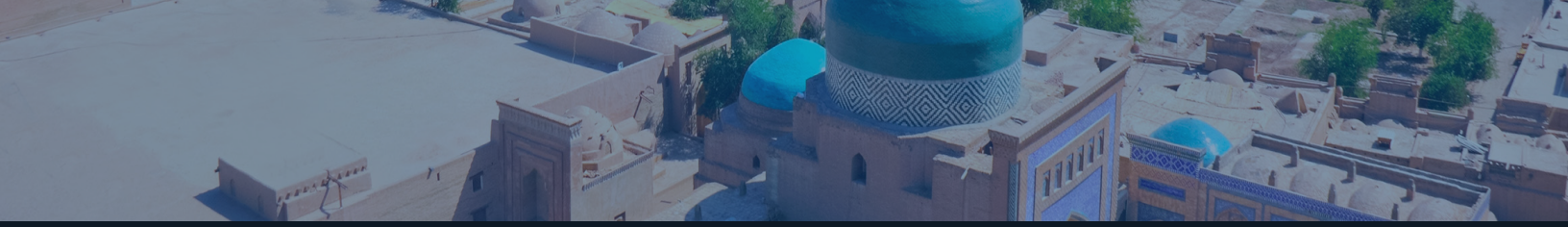
ClearPic compiles profile data for companies and people from diverse sources, including tax records, watchlists, data on public procurements, litigation data, and other public records.

The connected data is akin to a knowledge graph that reveals the structure of companies and the links between politically exposed persons (PEPs) and companies that may indicate corruption. For example, the relatives of a president might own coal mines that are receiving favorable treatment from government policies supported by the president.

Companies leverage ClearPic's data for performing due diligence on potential business partners, suppliers, and customers in Central Asian countries such as Uzbekistan, Tajikistan, Azerbaijan, Mongolia, Kyrgyzstan, and Turkey. There is an ongoing need to periodically re-screen partners. Are they going bankrupt or being sued? For institutions performing know your customer (KYC) due diligence, the ClearPic data becomes yet another source with which to screen potential business partners.

The Challenge

These public data sources often only provide company names without unique identifiers, such as tax IDs, making it difficult to match and link data across different sources. Furthermore, the matching and linking must be repeated as databases are updated or when new data is published. ClearPic needed to build entity



The amount of data to process is daunting. The Kazakhstan court database alone was 5 to 7 million data points.

resolution technology to analyze large volumes of data and identify patterns or similarities that suggest two or more records refer to the same entity. The technology needed to handle variations in spelling, punctuation of names, and take into account other identifying attributes, including type of business, place of business, and company directors.

ClearPic simplified and combined information from various sources into a single master list. This easy-to-use due diligence system allows customers to search and generate reports, and is continually updated with new data.

The languages of the records include Tajik (a variety of Persian), Azerbaijani, Uzbeki, Mongolian, and Turkish. Most of the records are written in Cyrillic-based languages, and the first step is an automated way to locate candidate record matches so that a human can confirm the entities are the same.

A major barrier for ClearPic was matching names and records to correlate data from different sources.

Difficulties of Matching Names from Central Asia

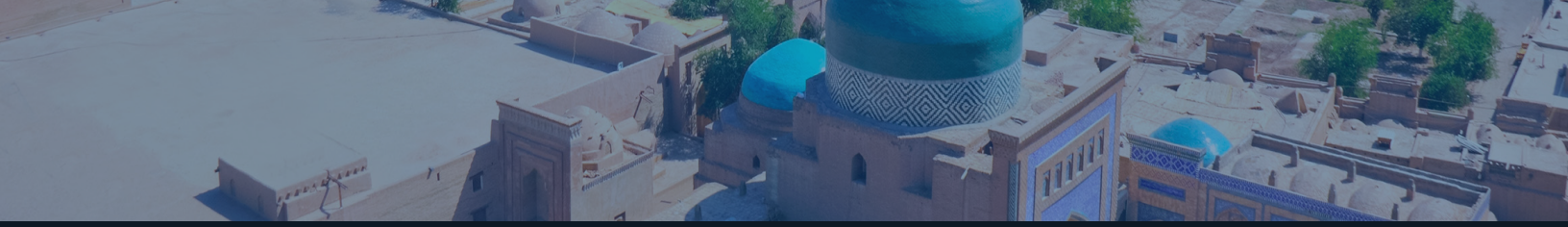
Before 2016, Tajik used the Russian structure of names (surname, name, paternal name), but beginning in 2016 the government started mandating the use of Tajik-style names (Iranian in nature).[2] Thus, the surname Rahmonov became written as Rahmon. Corporate records from 2014 to 2020 referring to the same person are therefore written completely differently!

The structure of Azerbaijani names resembles that of Turkish names and relies on patronymic constructs (son of/daughter of), such as "Ilham Haydar oghlu Aliyev" (where the patronym "Haydar oghlu" means "son of Haydar'").[3]

Uzbeki sources are written in the Uzbek language (in the Cyrillic or Latin script) and in the Russian language, so there are effectively three different names for the same person.

ClearPic is just starting to research Turkish and Mongolian names, which according to Artem Sentsov, ClearPic CTO "use a completely different naming structure [than the other three languages] and are very challenging." Mongolian also uses the Cyrillic script.

The amount of data to process is daunting. The Kazakhstan court database alone was 5 to 7 million data points. Other sources may only be about 100,000-500,000 records but are updated monthly without helpful change logs between updates, so ClearPic is also tracking historical changes. A company listing removed from a business registry is significant as it could indicate a firm went out of business or filed for bankruptcy.



> Connecting records is not just a 'one and done' integration, but an ongoing process, so the process must be repeatable and efficient.
>
> — Artem Sentsov, *ClearPic CTO*

"The data sources don't have an API or keep records in a clean nice way to query them and get data," Sentsov said. "Connecting records is not just a 'one and done' integration, but an ongoing process, so the process must be repeatable and efficient."

At first, ClearPic investigated building their own in-house matching tool.

"The solution was functional but faced issues with speed, efficiency, and quality," Sentsov said. "It resulted in numerous false positives and overlooked the accurate matches, which were the primary concern."

The Solution

The lack of precision and missed matches spurred Sentsov to start looking for a name matching tool from open-source providers and some well-known, commercial solutions. In the end he chose Babel Street Match. Why?

"Many of the solutions we considered matched records based on additional record fields like address and industry, but we often don't have anything except the name of the company or person," Sentsov said. "For us, Match functions as a one-stop shop solution for addressing the issue at hand."

Although Match does not explicitly support these five languages, "Match is able to get great [matching] results in these languages using just its Russian and Turkish language models," Sentsov said. "This was the killer feature of Match for us."

At the high end, ClearPic was finding 91-92% for precision (the number of correct matches) from Match, and 75-80% for recall (the total number of possible matches found).

With Match assisted integration of data from different sources, Sentsov and his team achieved good results using Match to deduplicate and match records. "Using Match tremendously increased our efficiency in aggregating and matching data," Sentsov said. "Previously, handling large amounts of data was a tedious task; however, with Match, the process has become more efficient and intelligent."

The Impact

ClearPic's customers are spread across industries, such as mining, finance, and brokers, but also legal firms and strategic consultants depend on ClearPic for performing their due diligence.

The tremendous advantage of ClearPic is that it enables customers to search for names in English and get results back from Cyrillic records — cleaned and aggregated from over 100 data sources that often do not contain ID numbers.

Now, ClearPic delivers clean, integrated public domain data to provide a visualization of ownership structure and the connection between companies to reduce risks when doing business in Central/Western Asia.

Endnotes

[1] United States Department of Justice, "Foreign Corrupt Practices Act," updated February 3, 2017. https://www.justice.gov/criminal-fraud/foreign-corrupt-practices-act

[2] "Tajikistan Bans Giving Babies Russian-Style Last Names," April 30, 2016, RadioFreeEurope RadioLiberty. https://www.rferl.org/a/tajikistan-bans-giving-babies-russian-style-last-names/27708093.html

"Tajik Lawmakers Approve Bill Banning Russified Surnames" April 29, 2020, RadioFreeEurope RadioLiberty. https://www.rferl.org/a/tajik-lawmakers-approve-bill-banning-russified-surnames/30583762.html

[3] https://en.wikipedia.org/wiki/Azerbaijani_name

For us, Match functions as a one-stop shop solution for addressing the issue at hand.

— Artem Sentsov, *ClearPic CTO*

Babel Street is the trusted technology partner for the world's most advanced identity intelligence and risk operations. The Babel Street Insights platform delivers advanced AI and data analytics solutions to close the Risk-Confidence Gap.

Babel Street provides unmatched, analysis-ready data regardless of language, proactive risk identification, 360-degree insights, high-speed automation, and seamless integration into existing systems. We empower government and commercial organizations to transform high-stakes identity and risk operations into a strategic advantage.

Learn more at **babelstreet.com.**

BABELSTREET