The 1st stage of openning up Government Data of the Republic of Kazakhstan

Yesmukhanov Dauren
Faculty of Information Technology
Kazakh-British Technical University
Almaty, Kazakhstan

Ross Horne
Faculty of Information Technology
Kazakh-British Technical University
Almaty, Kazakhstan

***Abstract*—Government data needs to be to posting online statistical information characterizing state and dynamics of the relevant industry, because the government transparency increases citizens' awareness of results of state activities and the interest of foreign investors in the national economy. This paper describes the way of increasing the transparency of national economy by integrating open government data related to public companies. Our approach uses interlinking datasets provided by different state and state-owned agencies with existed linked data source, available at egav.kz and well known as the first Linked Open Government Data of Kazakhstan. The result of the paper is more than 2 million triples, which describes ownership and financial information of companies significantly affecting to the national economy.**

Keywords—open data; open government data; linked data; triple store

I. *Intoduction*

Linked Data [1] is a methodology for publishing raw data on the Web using URIs to connect related datasets that were not previously interlinked. We use W3C standards such as RDF, OWL and SPARQL to join new datasets describing the national economy with the first Linked Open Government Data of the Republic of Kazakhstan.

Sovereign Wealth Fund “Samruk-Kazyna” is established in order to enhance competitiveness and sustainability of the national economy and prevent any potential negative impact of changes in the world markets on economic growth of the country. The company owns either in whole or in part, many nationally strategic companies in the country. It safeguards interests of the state as the sole shareholder of the company. The transparency of the fund's activities is essential to involve more people in the development of the national economy. On the one hand participating in the “National IPO’” program, which allows to citizen become co-owner of national companies, on the other hand the on-going privatization program stimulates small and medium business and invites foreign investors.

The related data are already available on the Web; however, the contribution of this paper is integrating of national economy data with the first Linked Open Government Data of Kazakhstan. Furthermore the integration allows executing custom query to obtain answers for question that cannot be posted elsewhere.

Section “Data Sources” lists all available raw open data sources published by “Samruk-Kazyna” and state agencies. These data sources were converted into a consistent, well-structured and organized dataset. Section “Challenges” highlights challenges faced during interlinking new datasets with existed linked data.

II. *Data Sources*

TABLE I. *DATA AND DESCRIBING ONTOLOGY*

Data	**Ontology class**
Executive body	http://data.egav.kz/ontology/publicCompany#Exectutives
Affiliates	http://data.egav.kz/ontology/publicCompany#Affiliates
Licenses	http://data.egav.kz/ontology/license\#
Board of directors	http://data.egav.kz/ontology/publicCompany#BoardOfDirectors
Shareholders	http://data.egav.kz/ontology/publicCompany#Shareholders
Balance sheet [2]	http://data.egav.kz/ontology/finance/report#BalanceSheet
Income statement [2]	http://data.egav.kz/ontology/finance/report#IncomeStatement
Cash flow statement [2]	http://data.egav.kz/ontology/finance/report#CashFlowStatement

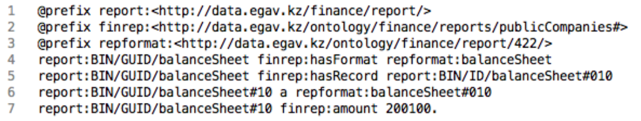
Table1 lists data scopes, which reflected in financial reports published by public company. The depositary of financial reports, available at *dfo.kz*, collects such documents and consolidates it manually. According the regulation of National Bank of the Republic of Kazakhstan [4] each public company must upload own financial reports periodically at *dfo.kz*.

At this stage I connected people (citizens or not) to the first Linked Open Government Data of the Republic of Kazakhstan. The following well-known ontologies are used as super classes and super properties:

- Companies and persons as Agents and Persons from FOAF [5];

- Executives, affiliates, board of directors and shareholders as Memberships from the organization ontology [6];

The ontology describers all properties and relations between entities, financial reports and records, but there are several official formats for report depend on the type of economic activity of public entity. The most of companies provides financial reports according International Financial Reporting Standards [3], but some companies prefer to use national standards. Each format is represented by code, according the regulation of National Bank of the Republic of Kazakhstan [4].



```
@prefix report:<http://data.egav.kz/finance/report/>
@prefix finrep:<http://data.egav.kz/ontology/finance/reports/publicCompanies#>
@prefix repformat:<http://data.egav.kz/ontology/finance/report/422/>
report:BIN/GUID/balanceSheet finrep:hasFormat repformat:balanceSheet
report:BIN/GUID/balanceSheet finrep:hasRecord report:BIN/ID/balanceSheet#010
report:BIN/GUID/balanceSheet#10 a repformat:balanceSheet#010
report:BIN/GUID/balanceSheet#10 finrep:amount 200100.
```

Fig. 1. *Example of pseode triples of financial report*

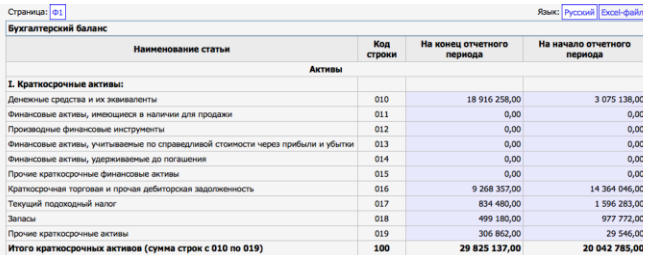
Four triples describe one financial report:

- URI *report/BIN/GUID/*, where BIN is business identification number of company provided by Tax Committee, GUID is generated unique identifier;
- Type *ontology/{format}/{section}#rowNum;*
- Publishing date *xsd:date;*
- Source *xsd:anyUri*.

Three triples describe one record from financial report:

- URI *report/BIN/GUID/{section}#rowNum;*
- Type *ontology/{format}/{section}#rowNum;*
- Value *xsd:integer|xsd:double|xsd:string.*

The depositary of financial reports is used as the major data source for collecting information above. Fig.2 shows an example of such web-page. But not all reports are already consolidated and data provided only in inconvenient formats like PDF and scanned JPEG. Therefore the portal of the Financial Supervision Agency, available at *afn.kz*, is used as additional source for graph of affiliated entities.



Бухгалтерский баланс

Наименование статьи	Код строки	На конец отчетного периода	На начало отчетного периода
Активы			
I. Краткосрочные активы:			
Денежные средства и их эквиваленты	010	18 916 258,00	3 075 138,00
Финансовые активы, имеющиеся в наличии для продажи	011	0,00	0,00
Производные финансовые инструменты	012	0,00	0,00
Финансовые активы, учитываемые по справедливой стоимости через прибыли и убытки	013	0,00	0,00
Финансовые активы, удерживаемые до погашения	014	0,00	0,00
Прочие краткосрочные финансовые активы	015	0,00	0,00
Краткосрочная торговая и прочая дебиторская задолженность	016	9 268 357,00	14 364 046,00
Текущий подоходный налог	017	834 480,00	1 596 283,00
Запасы	018	499 180,00	977 772,00
Прочие краткосрочные активы	019	306 862,00	29 546,00
Итого краткосрочных активов (сумма строк с 010 по 019)	**100**	**29 825 137,00**	**20 042 785,00**

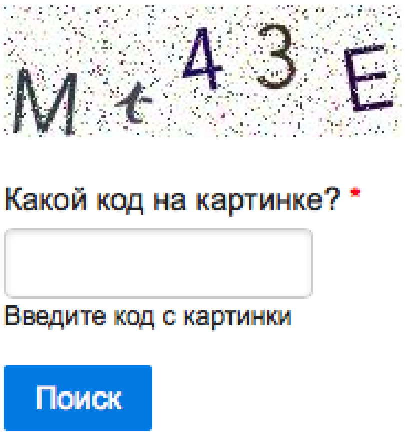
Fig. 2. *Example of pseode triples of financial report*

III. *Challenges*

Scraping Web page was solved by report and format specific scripts, row numbers are already provided at page for interlinking record type.

Resident affiliates were published without required identifier: BIN for legal entities and IIN (individual identification number) for private persons. To address one half of the problem, the Duplicate Detection Toolkit (DuDe) [7] supports searching for tuples that represent the same real world object in a variety of data sources. For instance, a non-standardized legal entity name can be matched with an official company name.

If legal entity name is quite unique key, surname and name of private person is not. However *afn.kz* provides birth date, so probability of exact matching is higher. The portal uses CAPTCHA for preventing crawlers, the example of such search page is shown in Fig.3. There are two solutions: human-based CAPTCHA solving [8] and optical character recognition (OCR). The fist solution promises high accuracy and low delays for recognition, but it requires additional expenses. This solution is cost effective, however low accuracy introduces a delay getting IIN of affiliated person. To deal with slow fetching, we have employed a RabbitMQ [9] based distributed system of crawlers and publishers, that runs scrapping scripts on multiple Amazon E2 instances.



Какой код на картинке? *

Введите код с картинки

Fig. 3. Example web form with CAPTCHA

There are difficulties with non-residents; generated GUID is used for URI matching.

Linking between data sources turns the first Linked Open Government Data from a collection of isolated datasets into a national data space. For each entity type, we used the following URI patterns consistently across multiple datasets:

- *http://data.egav.kz/legalEntity/{BIN}*
- *http://data.egav.kz/person/{IIN}*
- *http://data.egav.kz/report/{BIN}/{GUID}*
- *http://data.egav.kz/report/{BIN}/{GUID}/{section}*
- *http://data.egav.kz/report/{BIN}/{GUID}/{section}#RN*

- *http://data.egav.kz/license/{BIN}/{code}*

IV. *Economic impact*

Data and a SPARQL-endpoint are available at *egav.kz*. The following examples of queries expain benefits of using the first Linked Open Government of Kazakhstan:

1) *Total assets and liabilities of banks;*

2) *Changes in cash flow of state-owned companies and organizations;*

3) *The number of connection bewteen one company to another across affiliated entites and persons;*

4) *Find the most power shareholders in country accorind equity in balance sheet and share;*

5) *Find the most profitable sectors of economic;*

6) *Extend economic activity using information based on earned licenses.*

Open data about competitiveness in the local market can attract foreign investors and develop new business areas.

V. *Conbritbution*

The result of this paper is 2 millions of triples which describes the following:

- 20 453 financial reports;
- Executives and board of directors of 3502 public companies;
- Affiliated connections and shares;

All triples are interlinked with core data of the first Linked Open Government Data and support versioning based on named graphs.

VI. *Conclusion*

The first Linked Open Government Data is unique in the financial context of Kazakhstan and Central Asia. This is the only dataset that publishes open financial reports and relations as Linked Data. The dataset enables any citizen or foreign investor to obtain answers for question that cannot be posted elsewhere. Standardized ontologies enable complex analyses of finance and non-finance market of Kazakhstan.

This dataset is a bottom-up foundation for extending national Linked Data by state agencies, businesses and organizations, who wish to contribute their own data.

References

[1] Bizer C., Heath T., Berners-Lee T. Linked data-the story so far. – 2009.

[2] Williams, Jan R., et al. Financial and managerial accounting. China Machine Press, 2005.

[3] Tiffin R. International Financial Reporting Standards //London: Thorogood. – 2004.

[4] National Bank of the Republic of Kazakhstan . Rules publishing online financial statements, stock exchange information, corporate events and audit reports. Technical report, Ministry of Finance, Kazakhstan, 2012.

http://www.kase.kz/files/normative_base/post_72_240212.pdf

[5] Brickley D., Miller L. FOAF vocabulary specification 0.98 //Namespace document. – 2012. – T. 9.

[6] Reynolds D. An organization ontology //URL http://www. w3. org/TR/vocab-org. – 2010.

[7] Draisbach, Uwe, and Felix Naumann. "DuDe: The duplicate detection toolkit." Proceedings of the International Workshop on Quality in Databases (QDB). Vol. 100000. No. 1000000. 2010.

[8] Von Ahn L. et al. recaptcha: Human-based character recognition via web security measures //Science. – 2008. – T. 321. – №. 5895. – C. 1465-1468.

[9] Videla, Alvaro, and Jason JW Williams. RabbitMQ in action: distributed messaging for everyone. Manning, 2012.