A Framework for Applying Data Integration and Curation Pipelines to Support Integration of Migrants and Refugees in Europe

Oya Deniz Beyan[1], Siegfried Handschuh[2], Adamantios Koumpis[2(✉)], Garyfallos Fragidis[3], and Stefan Decker[1]

[1] RWTH Aachen University, Informatik 5, Aachen, Germany
{beyan,decker}@dbis.rwth-aachen.de
[2] Fakultät für Informatik und Mathematik, Universität Passau, Passau, Germany
siegfried.Handschuh@uni-passau.de,
adamantios.koumpis@uni-passau.de
[3] Institute of Service Science, University of Geneva, Geneva, Switzerland
gary.fragidis@gmail.com

Abstract. We investigate the benefit of data integration and curation services for the current refugee crisis and proposed an architecture to support development of innovative solutions. We focus on developing a multi- /cross-lingual semantic data curation pipeline enriched with natural language processing capabilities in order to (a) improve decision making capabilities of public authorities with data driven dashboards; (b) stimulate the development of innovative application and services supporting integration of refugees; and (c) improve the use of open data for tackling the societal challenges.

Keywords: Data curation of semi-structured and unstructured data · Big data analytics · Technology integration · Migration and refugee crisis · Social integration · Data lakes

1 Introduction

Open data initiatives and big data technologies can help EU economy to create better solutions for effective use of infrastructure such as housing and education, delivery of integration services for various domains including public health management and social cohesion, and developing a workforce to move towards a skill based economy [1, 2]. Moreover, mobile technologies and web based services can exploit the rich open data sources to create innovative products [3]. Despite of being a tremendous resource, open data, especially open government data, is yet largely untapped [9]. According to an assessment carried out by the European Data Portal team Open Data Maturity is just 44 % whereas Open Data readiness is 45 % in member states [10]. Although Open Data can help overcome certain economic constraints, developers tended benefit only small range of it, most limited with transportation and mobility.

H. Afsarmanesh et al. (Eds.): PRO-VE 2016, IFIP AICT 480, pp. 588–596, 2016.
DOI: 10.1007/978-3-319-45390-3_50

The major obstacle to develop innovative knowledge services and applications by exploiting the open data is related with the lack of the data integration and curation services [4]. Today, many service and mobile app developers requires access to timely and interpretable data to address raising societal needs and challenges. Most of the times they experience difficulties to identify data sources, understand the data representation schemas and accessing requirements, interpret different formats and merge into single usable schema, moreover deal with multilingual data [5]. All these preprocessing steps does not fit the real life requirements of entrepreneurs. In most of the cases, companies are under pressure to deliver their products in short time to answer raising demand. Moreover, their human capital and expertise in technologies are shaped around their core business, which not necessarily capable of dealing with data integration challenges. Therefore, next generation companies will need supporting data curation services to exploit the open data in their product development cycle.

In this research, we investigate the benefit of data integration and curation services for the current refugee crisis and proposed an architecture to support development of innovative solutions.

2 Refugee Crisis and Promises of Data Driven Services

The unprecedented human crisis generated by the wars in European Union neighborhood such as Libya, Syria, Iraq as well as further conflicts in Africa -Somalia or Middle East - Afghanistan have put a tremendous pressure on EU social and political system, with refugee waves never encountered since the Second World War. According to the International Organization for Migrants (IOM), since January 2015, 1.103.496 migrants, including asylum seekers, are reported to have arrived to Europe by land and sea routes [11]. Only in 2015 Germany alone received ~1.1M refugees [12]. The current humanitarian crisis is unprecedented with an appalling cost in short run, whereas in the long-run, much will depend on how well successful refugees are integrated [6].

The influx of refugees and migrants into Europe is conjuring new frontiers: a unique feature of most adult refugees is the connectivity of these refugees via smartphones that functions as their main lifeline to the wider world [7]. To this, several apps have been launched to help refugees integrate in recent months while activists have turned to crowdfunding and other online initiatives to help refugees find housing or jobs [13–15]. However, what may look, at first sight, like a lively ecosystem is only a collection of apps that have been developed independently and will disappear unexpectedly because of absence of use, leaving an empty space that will be difficult to be covered by a second generation of applications or services.

What we see as an unprecedented opportunity is in case we can come up with the necessary business model innovations that will help shift the balance from burden to benefit by making use of data driven decision making and service provision. In this scope our aim is to support independently developed applications and to strengthen the sustainability of the adhocracy of the existing ‘ecosystem’ by means of offering a systematized support for curating data.

From our side we consider that it is in the interest of all data users, i.e., not only the refugees, asylum seekers and migrants themselves but also the authorities in the several countries of EU as well as non-governmental organisations (NGOs) and activists to be able to access high-quality data which have integrity.

There are a set of challenges to produce and maintain high-quality data which goes beyond the capabilities of any independent app creator. First challenge is regarding accessing data required for smart policy making at governmental and administrative level. Although there are considerable amount of open data initiatives in EU, it is difficult to collect and curate data required planning for basic needs of refugees such as housing, employment and education. Next challenge is the variety of the data sources including websites, social media, open government portal provides wide range of structured, semi structured and unstructured data. Lastly cross lingual support is a challenge since the source and receiver of information requires different language sets.

3 Scope and Underlying Technologies

The main objective of our work is develop a multi-cross lingual semantic data curation pipeline enriched with natural language processing capabilities in order to (a) improve decision making capabilities of public authorities with data driven dashboards; (b) stimulate the development of innovative application and services supporting integration of refugees; and (c) improve the use of open data with semantic services for tackling the societal challenges. Such a data curation pipeline provides multilingual data integration services for data driven product development and will offer APIs to any end user that wants to connect and provide applications for the needs of migrants and refugees as well as for the need of governments and NGOs.

This poses the need for taking a more holistic approach to the entire data value chain that our research addresses, as described in the next section. In particular, our work concentrates on collecting, aggregating and integration data from a multitude of sources, which broadly fall into three categories:

- *Closed data sources:* private data which is available within the member organisations of the consortium such as refugees' personal health records [16]. This private data has to be protected against both accidental data leakages and unauthorized access.
- *Open data sources:* sources with data which has been made available to the public due to legislation such as the "freedom of information" act. This data mostly includes statistical data. Examples include World Health Organization's global health observatory data, European Union Open Data Portal, data sets from open health data, European Data Portal, World Bank Health Data, open government data portals [17–22].
- *Social media sites:* user generated content from social media sites has to be accessed and integrated on a per-site basis. Data from such sites will provide knowledge about refugees.

Most of the aforementioned data sources contain also geospatial data for which the data fusion process has more specialized requirements. Due to the nature of the data, the

geographical data points have to be combined based on geographical features such as e.g. proximity. In addition, the temporal nature of e.g. movement data needs to be preserved.

Linked Data and Semantic Web technologies simplify data integration for knowledge-intensive applications by enabling a Web of interoperable and machine-readable data based on formal and explicit descriptions of the structure and semantics of the data [23]. The fundamental building blocks for Linked Data are the graph-based data model of RDF [24], the Linked Data principles [25], and formal and domain specific semantics [26]. The benefits of Semantic Web technologies include simplification of information retrieval [27], information extraction [28] and data integration [29].

The Linked Data principles have been adopted by an increasing number of data providers, especially from the Linking Open Data community, which makes free and public data available as Linked Data. At the time of writing, approximately 300 different sources following the Linked Data principles are available. Specifically linked open government data promises an opportunity for people and companies to earn money and reap value from this high-quality and free information out there [8].

For many organizations who publishes open data there is a tradeoff between high quality data generation (requiring non-trivial human expert input) and massive data generation (requiring low human processing cost) [18]. As low cost open data publishing becomes more predominant, the urge for advanced data integration and curation approaches increases. In this sense, Linked Data and related Semantic technologies provide the foundational infrastructure to enable data fusion using data from different sources. Linked Data allows for addressing the structural, syntactic or semantic heterogeneity of data resulting from data access to multiple data sources with different formats, schemas or structure [30, 31].

4 Requirements and Design Principles

The proposed semantic data curation pipeline architecture aims to provide a service integration and collaboration platform for small and medium size enterprises to foster the development of data driven services and products. Today big data start to transforms businesses in all over the world and already many entrepreneurs created business models reliant on data. Although big data offers substantial value to organizations, it also brings many challenges to deal with data variety and quality issues. Addressing societal challenges with big data driven solutions may depend on consuming third party generated data created in diverse contexts with different set of goals. Adaptation and repurposing of the third party data requires specific expertise and in most cases involvement diverse technologies at the different stages of the data curation pipeline. The semantic data curation pipeline will enable emergence of high quality data driven business models for knowledge organizations. It will serve to companies, local and central governments and NGOs to build their data driven services and products to tackle social, economic and health care challenges of migrants and refugees. Smart application related with housing or job allocation, multi lingual screening programs related with emerging public health threats, targeted training and education programs to build qualified workforce can be mentioned among them.

The core value of the semantic data curation pipeline will be improving data quality and completeness as well as ensuring the verifiability and data provenance. Trustable and high quality data sets will have a positive impact on businesses by improving decision making capabilities based on big data analysis, It will also serve as a collaboration framework for small and medium size businesses to build their products in a value change, such as integrating cloud storage solution, multilingual translation services and data analysis solution from different partners within a single production line. In our work we have defined four guiding design principles for the semantic data curation pipeline to meet specified requirements of businesses:

(1) To facilitate the orchestration of mature and readily available technologies
The proposed semantic framework should support service integration for delivering high quality and verifiable data. It should allow for 3rd parties i.e. government organizations and NGOs to introduce, test and validate their Apps or systems as a part of the curation pipeline. Introduced technologies can support one or more curation level which comprises different roles as follows:
   (a) Discovery and acquisition of structured and unstructured data typically by registering to the Data Lakes (see also [32]); (b) Interpretation, annotation and providing multilingual support. Also includes indexing of collected, analyzed and curated data for efficient querying, metadata maintaining and semantic linking; (c) Analyzing data for posterity ensuring reliability, accessibility and retrieval capacity for future use and reuse; (d) Storing data for long term access, use and reuse; (e) providing access to external parties with a focus of selling integrated data and offering data driven services and products

(2) To improve the availability of multilingual machine discoverable data
Lack of multilingual machine discoverable data sets is one of the major shortcomings for stakeholders who are tackling challenges related with migration issues at large. The semantic data curation pipeline and big data platform should help to create large-scale, multilingual, semantically interoperable integrated data assets that will be of value and utility to the following perspectives:
   (a) *IT Environment Set-up:* By allowing connection to data lake API, to enable NGOs, Governments and European companies to build innovative multilingual products and services to help various groups of migrants and refugees understand the basics they can avail in each EU country. (b) Social objectives: to support refugees and migrants in their adaptation and integration, and facilitate public authorities and NGOs in optimizing their resources by data driven decision making.

(3) To create a multi sectoral value chain to address the needs of refugees
The value of the aspired infrastructure will be serving as a one-stop service platform to support needs of refugees and migrants. The value chain should be achieved by integrating and linking multi sectoral data by developing a contextual model to address social and economic needs of refugees. A contextual model for life events of targeted population can facilitate integrating information around needs in regard to immediate health risks, education, social support and employment. Our aim is to specify, develop and evaluate a life-event oriented, integrated, interoperable Pan-European platform for online one-stop analysis, filtering and visualization of added

value information and provision of services for refugees. Such a platform would be accompanied by a coherent services and products for realising and exploiting refugees' needs.

(4) To accelerate the development of new products and services
The curation pipeline can accelerate development of new products and innovative reuse of existing ones by supplying high quality curated data sets. The various use cases can demonstrate how a data management or data analytics product successfully marketed in another domain can benefit from data curation to extend its market. Products that applies the big data analytics solutions in various domains can be easily adopted to provide solutions for refugee crisis.

5 The Proposed Data Value Chain

We propose a five layer semantic data curation pipeline to meet above mentioned requirements and principles of value chain. It should be noted that the aforementioned requirements are not author-driven i.e. they were not dictated by the authors of this article but were acquired through the exploratory phase of our research to reflect real needs that have appeared either from the demand side namely the refugees themselves or the supply side namely the companies or the organisations that have offered Apps and services to satisfy such needs. The data value chain consists of 5 stages spanning from the data acquisition phase to the data usage phase. Layers and corresponding set of technologies is involved for extracting and integrating data from a variety of open data sources and delivering semantically enriched data to consumers as follows:

1. *Data Acquisition:* Technology responsible for discovering and registering structured and unstructured data from open data sources like Pan-European Open Data, Twitter, Facebook, etc. With this technology, an archiving platform will retrieve, structure and process large amounts of web content applying intelligent harvesting operations over hypertext, images, linked files, among other associated contents.
2. *Data Interpretation and Multilingual Interoperability:* Technology responsible to ensure multilingual translation of discovered and registered data from open data sources, building upon metadata management an indexing system for efficient querying and browsing over the data collected. Within this technology, proposed industry standard formats such as TMX will be applied for creation and maintenance of translation memories.
3. *Data Analysis and Curation:* Technology responsible for enabling the extraction of annotated datasets from structured and unstructured data, providing a semantic enrichment over the data which enhances the data usage and analysis. Within this technology a suite of statistical services will be provided as the tools to analyse, explore and extract insights based on prediction and correlation of data.
4. *Data Storage:* Technology responsible for the provision of a public-facing web repository, providing a scalable and secure storage system of the collected data, and an archiving mechanism, enabling an easy way to preserve, view, interrogate and reuse the content data.

5. *Data Usage:* This step comprises an API layer, localization service, and an analytics and decision making platform for data consumption.
   - *API engine:* This technology will be responsible for providing and managing accesses to external players consuming the curated data.
   - *Localization services:* This technology will be used to translate information into languages used by major migrant groups.
   - *Analytics and decision making platform:* This technology will be used for consuming the curated data and supporting employment and recruitment needs for refugees, migrants and asylum seekers, exploring information of various European countries and regions.

6 Conclusions

Currently, the use of publicly available data to facilitate interactions between public administrators, business and refugees is limited. Mostly the refugees use very specific applications developed in each country. Typical functionalities offered by existing applications allow refugees to access very particular information. Our proposed semantic curation pipeline will improve comprehension and increase accessibility of open data for developing services to tackle integration problems of refugees, thereby ensuring effective transparency of information. In order to accomplish this vision, we will realize an integrated five step technical solution so that at the end refugees will receive valuable information about their life-events. For this goal, we have developed a framework to provide a systematic transfer of information and technology across different sectors and a developed a data sharing and linking culture.

The benefits we see from the proposed system are related at a large extent to the benefits of any system that makes use of open public data in terms of fostering innovation. However, early adopters of open data driven solution should be aware of three possible pitfalls, namely: "(1) thinking of open data as a technical process rather than a culture change or innovation process; (2) getting trapped in outdated business models of data sale as revenue; (3) starting with the data rather than specific issues or challenges" [33]. In our case, we believe that targeting the societal challenge that Europe faces nowadays will lead the proposed framework in a sustainable innovation platform and open new opportunities for solving business and societal challenges.

References

1. Parycek, P., Hochtl, J., Ginner, M.: Open government data implementation evaluation. J. Theor. Appl. Electron. Commer. Res. **9**(2), 80–99 (2014)
2. Chen, H., Chiang, R.H.L., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
3. Lee, M., Almirall, E., Wareham, J.: Open data and civic apps: first-generation failures, second-generation improvements. Commun. ACM **59**(1), 82–89 (2015)

4. Lee, G., Kwak, Y.H.: An open government maturity model for social media-based public engagement. Gov. Inf. Q. **29**(4), 492–503 (2012)
5. Cavanillas, J.M., Curry, E., Wahlster, W.: New Horizons for a Data-Driven Economy, A Roadmap for Usage and Exploitation of Big Data in Europe. Springer, Heidelberg (2016)
6. OECD Migration Policy Debates, "Is this humanitarian migration crisis different?", N7, September 2015. https://www.oecd.org/migration/. Accessed: 22 Apr 2016
7. CBC News "For Syrian refugees, smartphones are a lifeline. http://www.cbc.ca/news/world/for-syrian-refugees-smartphones-are-a-lifeline-not-a-toy-1.3221349. Accessed 02 May 2016
8. Bizer, C., et al.: The meaningful use of big data: four perspectives–four challenges. ACM SIGMOD Rec. **40**(4), 56–60 (2012)
9. Ding, L., et al.: TWC data-gov corpus: incrementally generating linked government data from data. gov. In: Proceedings of the 19th International Conference on World wide web. ACM (2010)
10. Open Data Maturity in Europe 2015: Insights into the European state of play, European Union (2015)
11. World Migration Report 2015 – Migrants and Cities: New Partnerships to Manage Mobility, International Organization for Migration (2016)
12. 476.649 Asylanträge im Jahr 2015. https://www.bamf.de/SharedDocs/Meldungen/DE/2016/201610106-asylgeschaeftsstatistik-dezember.html. Accessed 24 Apr 2016
13. Ankommen: The guide for your first weeks in Germany. https://www.ankommenapp.de/
14. http://www.refugeespeaker.org/. Accessed 24 Apr 2016
15. http://www.universaldoctor.com/prod/en_GB/153/UniversalDoctor+Speaker+Web.html
16. European Commission Directorate-General for Health and Food Safety, Personal health record. http://ec.europa.eu/dgs/health_food-safety/docs/personal_health_record_english.pdf. Accessed 21 Apr 2016
17. Global Health Observatory, The data repository. http://www.who.int/gho/database/en/. Accessed: 27 Apr 2016
18. https://open-data.europa.eu/en/data
19. OpenDataHealth. http://openhealthdata.metajnl.com/. Accessed 22 Apr 2016
20. EuropeData Portal. https://publicdata.eu/. Accessed 22 Apr 2016
21. Health Data, The World Bank. http://data.worldbank.org/topic/health. Accessed 27 Apr 2016
22. Opening Up Government. https://data.gov.uk/. Accessed 30 Apr 2016
23. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
24. Decker, S., et al.: The semantic web: the roles of XML and RDF. IEEE Internet Comput. **4**(5), 63–73 (2000)
25. Berners-Lee, T.: Linked data-design issues. W3C (2006). Accessed 24 Apr 2016
26. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semant. Serv. Interoper. Web Appl. Emerg. Concepts **5**, 205–227 (2009)
27. van Elst, L., Dignum, V., Abecker, A.: Towards agent-mediated knowledge management. In: van Elst, L., Dignum, V., Abecker, A. (eds.) AMKM 2003. LNCS (LNAI), vol. 2926, pp. 1–30. Springer, Heidelberg (2004)
28. Fonseca, F.T., et al.: Using ontologies for integrated geographic information systems. Trans. GIS **6**(3), 231–257 (2002)
29. Fensel, D., et al.: OIL: an ontology infrastructure for the semantic web. IEEE Intell. Syst. **2**, 38–45 (2001)
30. Doan, A., Halevy, A.Y.: Semantic integration research in the database community: a brief survey. AI Mag. **26**(1), 83 (2005)
31. Moon, J.Y., Sproull, L.: Essence of distributed work: the case of the Linux Kernel. In: Hinds, P., Kiesler, S. (eds.) Distributed Work, pp. 381–404. The MIT Press, Cambridge (2002)

32. Stein, B., Morrison, A.: “The enterprise data lake: Better integration and deeper analytics”. Technology Forecast: Rethinking integration Issue 1. http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf. Accessed 4 May 2016
33. Carolan, L.: “The challenges and pitfalls from the use of open data”. Harvesting Open Data for Nutrition Security, ILSI Research Foundation, ILSI Annual Meeting 2016, St. Petersburg, Florida, 25 January 2016