PKGCubes: Personalizing Multidimensional Data Analytics through Personal Knowledge Graph Cubes

Fouad Zablith[1,*], Shadi Youssef[1]

[1]Olayan School of Business, American University of Beirut, PO Box 11-0236, Riad El Solh, 1107 2020, Beirut, Lebanon

Abstract

While knowledge graphs are increasingly adopted for supporting data analysis over linked data cubes, it is still challenging for end-users to personalize, preserve, and share cubes that are pertinent to their analytics objectives. Building on Personal Knowledge Graphs, this study introduces the notion of Personal Knowledge Graph Cubes (PKGCubes). PKGCubes serve as a mediator between the web of data cubes, and data analysis platforms. A demo of PKGCubes Manager is presented, enabling data analysts to create, publish, and reuse PKGCubes in standalone data analysis tools. This work contributes to offering more personalized and self-service analytics tasks on the growing web of data.

Keywords

Personal knowledge graph cubes, linked data, visual analytics, semantic web, OLAP, data cubes

1. Introduction

Increased research efforts are aiming to leverage the expressive nature of knowledge graphs for facilitating data analytics tasks [1]. One popular type of data is multidimensional data having measures and dimensions that form cubes for Online Analytical Processing (OLAP) [2]. In this context, related works ranged from studying the effective representation of data cubes through ontologies (e.g., the RDF Data Cube [3] and QB4OLAP [4]), to increasing the usability and value of the graph data [5] through visual [6] and knowledge graph management features [7].

While such efforts are providing greater data sharing and usability opportunities, manipulating knowledge graphs for data analytics still poses some challenges to end-users [8]. With the plethora of published linked open datasets, end-users find it challenging to customize, preserve, and share knowledge graph cubes that are pertinent to their analytical objectives. This demo paper focuses on answering the following research question: how can we better personalize data analysis over multidimensional web of data cubes?

2. Personal Knowledge Graph Cubes

Personal Knowledge Graphs (PKG) allow the representation of knowledge graph entities that are relevant and of personal nature to a particular individual [9]. We see an opportunity to build

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

*Corresponding author.

fouad.zablith@aub.edu.lb (F. Zablith); say09@mail.aub.edu (S. Youssef)
https://fouad.zablith.org/ (F. Zablith)
0000-0002-8978-9911 (F. Zablith)

on the notion of PKGs to enable a more personalized depiction of linked data cubes. We propose Personal Knowledge Graph Cubes (PKGCubes) to represent data cubes that fulfill an individual's analytical needs and requirements. Building on the rich knowledge graph semantics, PKGCubes are meant to be stored, shared, and combined with other cubes.

Figure 1 illustrates how we envision the PKGCube. It acts as a mediator between published linked data cubes and data analytics tools. It serves as a personal and persistent snapshot view of graph data, representing entities that connect to linked data cube sources, and feeds the personal data of interest into analytics tools. Ontologically, we design a PKGCube as an extension of the RDF Data Cube vocabulary [3]. While the RDF Data Cube vocabulary is well positioned to represent the data cube entities (e.g., observations, slices, etc.), it lacks the representation of entities needed to make them more personalized. This is the gap that PKGCubes aim to fill. In its initial ontology version, a PKGCube represents a: *person* entity who published the cube; *version* for tracking changes; *cube hash* to encode the content; *access* specification to set private versus public cubes; *source query* that enables its recreation; *description* with information on the data in the cube; *link* to where the data is available; *observations* derived from linked data cube sources; and *slice* information to store filtering settings applied to the PKGCube.

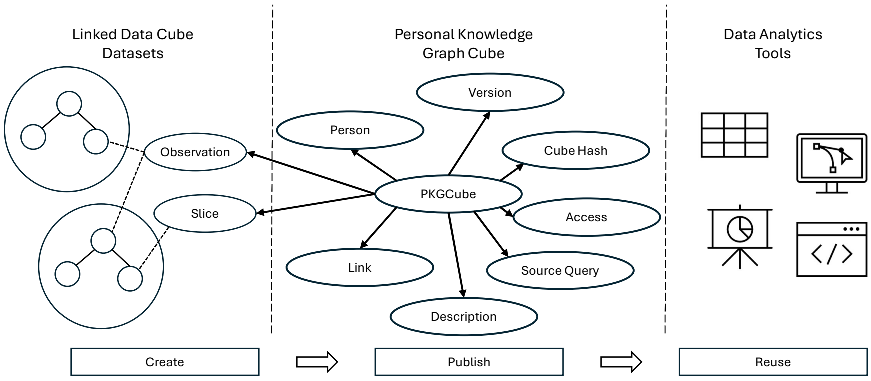


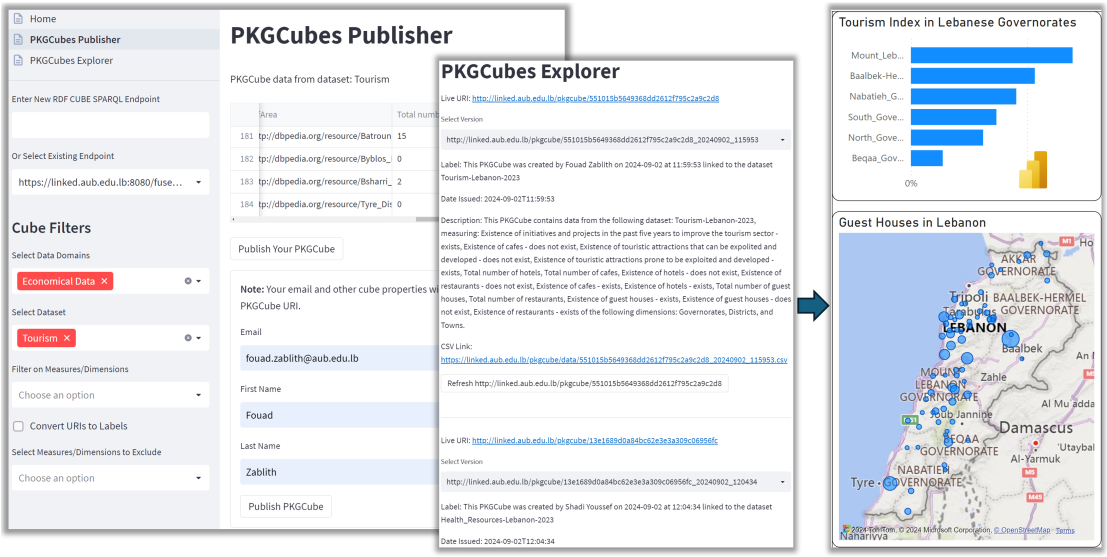
Figure 1: Illustration of Personal Knowledge Graph Cube Entities and the Related Framework Steps.

We envisage a framework to create, publish, and reuse PKGCubes. The creation of the PKGCubes involves providing individuals access to navigate and select entities from linked data cube sources to include in their PKGCube based on their individual analysis objectives. The created PKGCubes are stored and published on a triplestore. To maximize reusability of the PKGCubes in a variety of data analytics tools, they can be further processed and transformed into more manipulable data formats such as tabular structure in the form of spreadsheets or Comma Separated Values (CSV).

3. Demo: PKGCubes Manager

We demonstrate the feasibility of PKGcubes through *PKGCubes Manager*, a Python Streamlit online app[1] that enables data analysts to *create* PKGCubes through a set of filters, *publish* the

[1] PKGCubes Manager is accessible at: https://linked.aub.edu.lb:8502/.



Enter New RDF CUBE SPARQL Endpoint

Or Select Existing Endpoint

https://linked.aub.edu.lb:8080/fuse…

Cube Filters

Select Data Domains

Economical Data ✕

Select Dataset

Tourism ✕

Filter on Measures/Dimensions

Choose an option

☐ Convert URIs to Labels

Select Measures/Dimensions to Exclude

Choose an option

PKGCubes Publisher

PKGCube data from dataset: Tourism

	fArea	Total numb
181	tp://dbpedia.org/resource/Batroun	15
182	tp://dbpedia.org/resource/Byblos_	0
183	tp://dbpedia.org/resource/Bsharri_	2
184	tp://dbpedia.org/resource/Tyre_Dis	0

Publish Your PKGCube

Note: Your email and other cube properties wi
PKGCube URI.

Email: fouad.zablith@aub.edu.lb

First Name: Fouad

Last Name: Zablith

Publish PKGCube

PKGCubes Explorer

Live URI: http://linked.aub.edu.lb/pkgcube/551015b5649368dd2612f795c2a9c2d8

Select Version

http://linked.aub.edu.lb/pkgcube/551015b5649368dd2612f795c2a9c2d8_20240902_115953

Label: This PKGCube was created by Fouad Zablith on 2024-09-02 at 11:59:53 linked to the dataset Tourism-Lebanon-2023

Date Issued: 2024-09-02T11:59:53

Description: This PKGCube contains data from the following dataset: Tourism-Lebanon-2023, measuring: Existence of initiatives and projects in the past five years to improve the tourism sector - exists, Existence of cafes - does not exist, Existence of touristic attractions that can be expolited and developed - does not exist, Existence of touristic attractions prone to be exploited and developed - exists, Total number of hotels, Total number of cafes, Existence of hotels - does not exist, Existence of restaurants - does not exist, Existence of cafes - exists, Existence of hotels - exists, Total number of guest houses, Total number of restaurants, Existence of guest houses - exists, Existence of guest houses - does not exist, Existence of restaurants - exists of the following dimensions: Governorates, Districts, and Towns.

CSV Link: https://linked.aub.edu.lb/pkgcube/data/551015b5649368dd2612f795c2a9c2d8_20240902_115953.csv

Refresh http://linked.aub.edu.lb/pkgcube/551015b5649368dd2612f795c2a9c2d8

Live URI: http://linked.aub.edu.lb/pkgcube/13e1689d0a84bc62e3e3a309c06956fc

Select Version

http://linked.aub.edu.lb/pkgcube/13e1689d0a84bc62e3e3a309c06956fc_20240902_120434

Label: This PKGCube was created by Shadi Youssef on 2024-09-02 at 12:04:34 linked to the dataset Health_Resources-Lebanon-2023

Date Issued: 2024-09-02T12:04:34

Figure 2: Screenshots of the PKGCubes Manager Features.

cubes to a triple store, and *reuse* them in analytics apps such as Microsoft Power BI [10]. Figure 2 shows the main features of the PKGCubes Manager app[2]. We test the app in the context of openly accessible statistical data in several domains (e.g., health care, tourism, and others) that were transformed from various distributed data sources (e.g., ministries) in Lebanon. PKGCubes Manager has so far two main functionalities, the PKGCubes Publisher, and PKGCubes Explorer.

The *PKGCubes Publisher* enables data analysts to specify a SPARQL endpoint that contains RDF data cubes. It offers predefined endpoints in the drop-down menu, or new endpoints that can be provided by analysts. The selected endpoints need to store data cubes with explicit datasets, measures, and dimensions following the RDF Data Cube vocabulary. The publisher app scans the data available in the endpoint using SPARQL templates designed to detect the available datasets and their related entities. The SPARQL results are then used to populate the “Cube Filters” available in the app. Users can then filter the cubes based on the domains, datasets, and the available dataset measures and dimensions. After the selection, the tool builds a SPARQL query in the background based on the filters selection and presents the SPARQL results in tabular format. Users can then check the cube data loaded in the table, fine-tune the filters if needed, and publish the cube.

To publish the cube, users need to provide their personal details including their name and email. Then the app executes (1) a Unique Resource Identifiers (URIs) and linkage generation step, (2) a versioning check, followed by (3) a data publication phase. In the first step, the PKGCubes URIs are generated based on the MD5 hash of the following combination: <email+dataset+measures+dimensions>. This configuration enables associating a unique one-way identification of the PKGCubes while preserving the users’ data privacy. It also helps with storing the configuration that the user followed to generate the PKGCube, and appropriate

[2] A video demonstration is available at: https://youtu.be/e9NPsrVSXXM

linkages among cube versions. The publisher links the PKGCube URIs to the relevant entities (e.g., observations extracted from the endpoint, source query, and other elements mentioned in Figure 1) and the personal user URI generated based on the MD5 hash of their provided email. In the second step, a versioning functionality was implemented to keep track of the different versions of the same PKGCube. Versioning is valuable to have snapshots of the data saved at various points in time. The publisher app checks the version of the PKGCube at two levels. At the first level, the app checks whether the PKGCube URI was already published. If it's a new PKGCube, the PKGCube entities and related files (i.e., CSV and RDF) are generated. If the cube exists, it checks the extracted content from the cube, compares it to the cube content hash of the latest version, and creates appropriate version linkages that users can explore. Finally, the generated PKGCube entities are published to a triplestore.

In the *PKGCubes Explorer* part, data analysts are able to browse the published PKGCubes details, their linked versions, and reuse the related data in external applications. Another feature of the tool is a "refresh" functionality that updates the PKGCube with the latest data available in the initial endpoint. This is useful to handling cases when the source RDF Data Cubes content changes, allowing analysts to update their PKGCubes with the latest data that can be seamlessly reflected in their external applications. To illustrate the reuse of data, Figure 2 showcases how a PKGCube's linked CSV file was used in Microsoft Power BI to generate a dashboard on tourism index and guest houses around Lebanon[3]. This demonstrates the potential of PKGCubes to create personalized, uniquely referenced, and preserved data cubes that can be reused by analysts in their preferred data analysis environments.

4. Conclusion

We presented in this paper the notion of Personal Knowledge Graph Cubes, with a demonstration of its application through the PKGCubes Manager online app. As part of future research, this work can benefit from developing more robust management and access control features of PKGCubes. This conforms with the Personal Knowledge Graph ecosystem laid out by Skjaeveland et al. [11]. Another interesting research direction would be to investigate additional social interactions around the cubes. A possible approach to investigate is the potential alignment with the Social Linked Data (Solid) principles [12], which provide further privacy and user-control functionalities when publishing data [13]. We are planning to evaluate the impact of PKGCubes on performing data analytics tasks in projects and use cases. Use case data will help improve the ontology and interface design for managing PKGCubes. This research contributes to providing more personalized and self-service analytics [14, 15] on the growing web of data.

Acknowledgments

This work was partially supported by the Olayan School of Business (OSB) Research Initiative fund, and the American University of Beirut Research Board (URB).

[3] The PKGCube used to generate the Power BI visualizations is accessible at: http://linked.aub.edu.lb/pkgcube/551015b5649368dd2612f795c2a9c2d8

References

[1] M. E. Papadaki, Y. Tzitzikas, M. Mountantonakis, A Brief Survey of Methods for Analytics over RDF Knowledge Graphs, Analytics 2 (2023) 55–74. doi:10.3390/analytics2010004, number: 1 Publisher: MDPI.

[2] A. Abelló, O. Romero, T. B. Pedersen, R. Berlanga, V. Nebot, M. J. Aramburu, A. Simitsis, Using semantic web technologies for exploratory OLAP: a survey, IEEE transactions on knowledge and data engineering 27 (2014) 571–588. Publisher: IEEE.

[3] The RDF Data Cube Vocabulary, 2014. URL: https://www.w3.org/TR/vocab-data-cube/.

[4] L. Etcheverry, A. A. Vaisman, QB4OLAP: a new vocabulary for OLAP cubes on the semantic web, in: Proceedings of the Third International Conference on Consuming Linked Data, volume 905, CEUR-WS. org, 2012, pp. 27–38.

[5] P. Escobar, G. Candela, J. Trujillo, M. Marco-Such, J. Peral, Adding value to Linked Open Data using a multidimensional model approach based on the RDF Data Cube vocabulary, Computer Standards & Interfaces 68 (2020) 103378. doi:10.1016/j.csi.2019.103378.

[6] G. Tschinkel, E. E. Veas, B. Mutlu, V. Sabol, Using Semantics for Interactive Visual Analysis of Linked Open Data., in: ISWC (Posters & Demos), Citeseer, 2014, pp. 133–136.

[7] P. Haase, D. M. Herzig, A. Kozlov, A. Nikolov, J. Trame, metaphactory: A platform for knowledge graph management, Semantic Web 10 (2019) 1109–1125. Publisher: IOS Press.

[8] S. Ferré, Analytical Queries on Vanilla RDF Graphs with a Guided Query Builder Approach, in: T. Andreasen, G. De Tré, J. Kacprzyk, H. Legind Larsen, G. Bordogna, S. Zadrożny (Eds.), Flexible Query Answering Systems, Springer International Publishing, Cham, 2021, pp. 41–53. doi:10.1007/978-3-030-86967-0_4.

[9] K. Balog, T. Kenter, Personal knowledge graphs: A research agenda, in: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, 2019.

[10] Power BI - Data Visualization | Microsoft Power Platform, 2024. URL: https://www.microsoft.com/en-us/power-platform/products/power-bi.

[11] M. G. Skjæveland, K. Balog, N. Bernard, W. Łajewska, T. Linjordet, An ecosystem for personal knowledge graphs: A survey and research roadmap, AI Open 5 (2024) 55–69. doi:10.1016/j.aiopen.2024.01.003.

[12] A. V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Aboulnaga, T. Berners-Lee, Solid: a platform for decentralized social applications based on linked data, Technical Report, MIT CSAIL & Qatar Computing Research Inst., 2016.

[13] S. Meckler, R. Dorsch, D. Henselmann, A. Harth, The Web and Linked Data as a Solid Foundation for Dataspaces, in: Companion Proceedings of the ACM Web Conference, WWW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1440–1446. doi:10.1145/3543873.3587616.

[14] A. Abelló, J. Darmont, L. Etcheverry, M. Golfarelli, J.-N. Mazón, F. Naumann, T. Pedersen, S. B. Rizzi, J. Trujillo, P. Vassiliadis, Fusion cubes: Towards self-service business intelligence, International Journal of Data Warehousing and Mining (IJDWM) 9 (2013) 66–88. Publisher: IGI Global.

[15] J. Passlick, L. Grützner, M. Schulz, M. H. Breitner, Self-service business intelligence and analytics application scenarios: A taxonomy for differentiation, Information Systems and e-Business Management 21 (2023) 159–191. doi:10.1007/s10257-022-00574-3.