Extended RDF support for Biomedical Knowledge Graphs in pyBioDataFuse: on-the-fly RDF graph generation and new resource annotators

Javier Millán Acosta[1], Egon Willighagen[1], Yojana Gadiya[2,3] and Tooba Abbassi-Daloii[1]

[1]Department of Translational Genomics, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands

[2]Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg, Germany

[3]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn, Germany

Abstract

Abstract Integrating high-dimensional omics data with curated knowledge from publicly available databases provides crucial biological context to unlock systems biology analyses and interpret complex biological mechanisms. pyBioDataFuse is an open-source Python package designed to facilitate interoperability and integration complexities in omics workflows by querying for curated knowledge for genes and compounds across several databases, and wrangling the results in a harmonized format used to create of context-specific knowledge graphs (KGs). Moreover, pyBioDataFuse supports graph-based analyses with plugins for Neo4j and Cytoscape. However, there is an ever-growing set of resources that can be plugged into pyBioDataFuse via annotators, and the support of the downstream conversion of the generated property graphs into Resource Description Framework (RDF) graphs has been a pending task. We have addressed these issues by developing new annotators and a BioDataFuse Ontology and RDF module that allow to convert the resulting KGs into an RDF graph. The package supports the automated generation of Shape Expressions (ShEx) and Shapes Constraint Language (SHACL) shape graphs specific to the generated graph to support its validation and documentation. The ontology expands on existing resources to accurately represent the variety of scores, biological entities, processes and functions captured in a BioDataFuse graph. Altogether, we propose pyBioDataFuse as a FAIR tool for knowledge graph development and analysis.

Keywords

Keywords Biomedical Knowledge Graph, RDF, Context-Specific Knowledge Graph, Data Wrangling, Graph Analysis, Python

1. Introduction

BioDataFuse is a tool aiming to simplify the process of building a context-specific knowledge graph for datasets containing gene or compound rows. With pyBioDatafuse (the core Python package) providing most of the functionalities, the user has control over graph creation by selecting which kinds of annotations or curated knowledge and from which source to query for, and the results are automatically converted into a property graph, using NetworkX; or an RDF graph, leveraging the rdflib Python library.

While earlier iterations of BioDataFuse focused on prop- erty graph creation and on its integration with tools like Cytoscape and Neo4j, they lacked support for RDF graph conversion of the queried (meta)data. This limitation restricted interoperability and adoption in workflows relying on semantic web standards [1].

2. Annotated resources

BioDataFuse connects data from a range of curated resources to build knowledge graphs tailored to specific research needs. These resources can be grouped by the type of information they provide and

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025

0000-0002-4166-7093 (J. Millán Acosta); 0000-0001-7542-0286 (E. Willighagen); 0000-0002-7683-0452 (Y. Gadiya); 0000-0002-4904-3269 (T. Abbassi-Daloii)

the biological entities they describe:

- Identifier mapping and harmonization: BridgeDb Webservice [2].
- Gene expression data: queried on Bgee [3, 4], which provides detailed gene expression data on specific tissues and species to support comparative studies.
- Disease links: links between genes and compounds with disease or other phenotype data are queried on DisGeNET [5] (gene-disease associations) and Open Targets [6] (disease-compound relationships). AOP- Wiki RDF [7] is used to link adverse outcomes and key events tied to genes and compounds, aiding toxicology and risk analysis.
- Pathway data are queried on MINERVA [8], WikiPathways [9], and Reactome [10] (the last one via Open-Targets), which provide molecular interaction networks and curated pathways. Gene Ontology annotations [11], accessed through Open Targets, add information about biological processes, molecular functions, and cellular compartments.
- Molecular interactions are queried on Open Targets for linking genes and compounds, and PubChem assays for compound-protein interactions [12]. MolMeDB [13] focuses on membrane transporter interactions with compounds, and StringDB [14] offers data on protein-protein interactions.

The different annotators and SPARQL queries used to query databases are available on GitHub at pyBioDataFuse/annotators.

3. RDF Schema and Shapes

The complete content of the tables combining query results can be exported in RDF. The RDF schema for BioDataFuse graphs reflects the data types listed above and uses classes and predicates in the BioDataFuse ontology to provide the semantic layer for BioDataFuse data. Although there is just one overarching schema for the RDF version of the graphs, each BioDataFuse RDF graph contains nodes only for the specific entities involved in the queried databases. This is reflected in the generated shape graphs.

The package uses shexer [15] to generate the ShEx [16] and SHACL [17] shapes that can be used to document, validate and describe the resulting graph. The shape graphs are generated automatically from triples frequencies, but since they are stored as rdflib Conjunctive Graphs they can be fine-tuned at any point during workflows.

4. Future work

pyBioDataFuse is actively being developed on GitHub, and the latest version 1.0.0 is available on PyPI. Future development will focus on enhancing the annotator template, enabling the semi-automated generation of new annotators. Recent advancements highlight the use of VoID headers and shape extraction tools such as VoID generation, RDF-Config, SheXer, and others, for efficiently generating APIs, SPARQL queries, and other interfaces to access SPARQL data. BioDataFuse plans to incorporate these methods to facilitate the semi-automated creation of annotator templates.

Acknowledgments

We wish to acknowledge all the contributors to the pyBioDatafuse repository (see on GitHub pyBioDataFuse/contributors for a complete list).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Y. Gadiya, A. Ammar, E. Willighagen, D. Martinat, A. C. Sima, H. Balci, T. Abbassi-Daloii, BioHackEU23 report: Extending interoperability of experimental data using modular queries across biomedical resources, BioHackrXiv (2023).

[2] M. P. van Iersel, A. R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B. R. Conklin, C. T. Evelo, The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services, BMC Bioinformatics 11 (2010) 5.

[3] F. B. Bastian, J. Roux, A. Niknejad, A. Comte, S. S. Fonseca Costa, T. M. de Farias, S. Moretti, G. Parmentier, V. R. de Laval, M. Rosikiewicz, J. Wollbrett, A. Echchiki, A. Escoriza, W. H. Gharib, M. Gonzales-Porta, Y. Jarosz, B. Laurenczy, P. Moret, E. Person, P. Roelli, K. Sanjeev, M. Seppey, M. Robinson-Rechavi, The bgee suite: integrated curated expression atlas and comparative transcriptomics in animals, Nucleic Acids Res. 49 (2021) D831–D847.

[4] F. B. Bastian, A. B. Cammarata, S. Carsanaro, H. Detering, W.-T. Huang, S. Joye, A. Niknejad, M. Nyamari, T. Mendes de Farias, S. Moretti, M. Tzivanopoulou, J. Wollbrett, M. Robinson-Rechavi, Bgee in 2024: focus on curated single-cell RNA-seq datasets, and query tools, Nucleic Acids Res. 53 (2025) D878–D885.

[5] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Res. 48 (2020) D845–D855.

[6] D. Ochoa, A. Hercules, M. Carmona, D. Suveges, J. Baker, C. Malangone, I. Lopez, A. Miranda, C. Cruz-Castillo, L. Fumis, M. Bernal-Llinares, K. Tsukanov, H. Cornu, K. Tsirigos, O. Razuvayevskaya, A. Buniello, J. Schwartzentruber, M. Karim, B. Ariano, R. E. Martinez Osorio, J. Ferrer, X. Ge, S. Machlitt-Northen, A. Gonzalez-Uriarte, S. Saha, S. Tirunagari, C. Mehta, J. M. Roldán-Romero, S. Horswell, S. Young, M. Ghoussaini, D. G. Hulcoop, I. Dunham, E. M. McDonagh, The next-generation open targets platform: reimagined, redesigned, rebuilt, Nucleic Acids Res. 51 (2023) D1353–D1359.

[7] M. Martens, C. T. Evelo, E. L. Willighagen, Providing adverse outcome pathways from the AOP-Wiki in a semantic web format to increase usability and accessibility of the content, Appl. In Vitro Toxicol. 8 (2022) 2–13.

[8] P. Gawron, E. Smula, R. Schneider, M. Ostaszewski, Exploration and comparison of molecular mechanisms across diseases using MINERVA net, Protein Sci. 32 (2023) e4565.

[9] A. Agrawal, H. Balcı, K. Hanspers, S. L. Coort, M. Martens, D. N. Slenter, F. Ehrhart, D. Digles, A. Waagmeester, I. Wassink, T. Abbassi-Daloii, E. N. Lopes, A. Iyer, J. M. Acosta, L. G. Willighagen, K. Nishida, A. Riutta, H. Basaric, C. T. Evelo, E. L. Willighagen, M. Kutmon, A. R. Pico, WikiPathways 2024: next generation pathway database, Nucleic Acids Res. 52 (2024) D679–D689.

[10] M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, R. Haw, B. Jassal, L. Matthews, B. May, R. Petryszak, E. Ragueneau, K. Rothfels, C. Sevilla, V. Shamovsky, R. Stephan, K. Tiwari, T. Varusai, J. Weiser, A. Wright, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase 2024, Nucleic Acids Res. 52 (2024) D672–D678.

[11] Gene Ontology Consortium, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. Dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu,

R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J. Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, M. Westerfield, The gene ontology knowledgebase in 2023, Genetics 224 (2023).

[12] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, PubChem 2025 update, Nucleic Acids Res. 53 (2025) D1516–D1525.

[13] J. Juračka, M. Šrejber, M. Melíková, V. Bazgier, K. Berka, MolMeDB: Molecules on membranes database, Database (Oxford) 2019 (2019).

[14] E. Prud'hommeaux, J. E. Labra Gayo, H. Solbrig, Shape expressions, in: Proceedings of the 10th International Conference on Semantic Systems, ACM, New York, NY, USA, 2014.

[15] D. Szklarczyk, K. Nastou, M. Koutrouli, R. Kirsch, F. Mehryary, R. Hachilif, D. Hu, M. E. Peluso, Q. Huang, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, C. von Mering, The STRING database in 2025: protein networks with directionality of regulation, Nucleic Acids Res. 53 (2025) D730–D737.

[16] D. Fernandez-Álvarez, J. E. Labra-Gayo, D. Gayo-Avello, Automatic extraction of shapes using shexer, Knowl. Based Syst. 238 (2022) 107975.

[17] H. Knublauch, D. Kontokostas, Shapes constraint language (SHACL) W3C recommendation, 2017. URL: https://www.w3.org/TR/2017/REC-shacl-20170720/.