Making expert curated knowledge graphs FAIR

Jerven Bolleman[1], Alan Bridge[1,*], Nicole Redaschi[1] and UniProt Consortium[1,2,3,4]

[1]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, CMU, 1 Michel Servet, 1211 Geneva 4, Switzerland
[2]European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK
[3]Protein Information Resource (PIR), Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, USA
[4]Protein Information Resource (PIR), University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

Abstract

To address the users' need to combine knowledge from different expert curated resources, the biocuration community is heavily invested in the standardization of knowledge with shared ontologies and, more recently, in the representation of data in the form of knowledge graphs (KGs). To easily integrate data from, or query data across, different KGs it is necessary to also standardize the form in which they are published. The W3C standards RDF/OWL and SPARQL were created to address this need and enable the creation of a Semantic Web. Here we describe the use of these standards to publish public SPARQL endpoints for resources such as UniProt, Rhea and SwissLipids and RDF allowing private SPARQL endpoints on premise and in clouds (e.g. AWS, Oracle). At more than 110 distinct billion triples - RDF statements - UniProt is the largest freely available KG. UniProt and other SPARQL endpoints support complex analytical queries and inferences that go beyond queries through graph-based machine learning and other approaches. They integrate - federate - expert curated knowledge of protein function with biological and biochemical data from other KGs available in RDF or OWL like the Gene Ontology (functions), Bgee (expression patterns), OMA (orthology), and IDSM (chemical structures). They also serve as APIs to enhance website data display and data mining capabilities - for example to select and enrich SwissBioPics images to visualize subcellular location data, or to perform chemical similarity and chemical substructure search with IDSM directly in Rhea.

Keywords

protein, UniProt, SPARQL, Rhea, SwissLipds, lipids, reaction

1. Funding & acknowledgements

This work was supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), National Institute of Mental Health (NIMH), and National Cancer Institute (NCI) of the National Institutes of Health (NIH) under grant U24HG007822. Additional support for the EMBL-EBI's involvement in UniProt comes from European Molecular Biology Labo-

SWAT4HCLS 2023
*Corresponding author.
jerven.bolleman@sib.swiss (J. Bolleman); alan.bridge@sib.swiss (A. Bridge); nicole.redaschi@sib.swiss (N. Redaschi)
0000-0002-7449-1266 (J. Bolleman); 0000-0003-2148-9135 (A. Bridge); 0000-0001-8890-2268 (N. Redaschi)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3415/paper-17.pdf

ratory (EMBL) core funds, the Alzheimer's Research UK (ARUK) grant ARUK-NAS2017A-1, the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/T010541/1] and Open Targets. UniProt activities at the SIB are additionally supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI. PIR's UniProt activities are also supported by the NIH grants R01GM080646, G08LM010720, and P20GM103446, and the National Science Foundation (NSF) grant DBI-1062520.