Data in use for Alzheimer disease study: combining gene expression, orthology, bioresource and disease datasets

Tarcisio Mendes de Farias[1,2,*], Tatsuya Kushida[3,*], Ana C. Sima[1,2], Christophe Dessimoz[1,2], Hirokazu Chiba[4], Frédéric Bastian[1,2,†] and Hiroshi Masuya[3,†]

[1]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
[2]University of Lausanne, Lausanne, Switzerland
[3]BioResource Research Center, RIKEN, Tsukuba-shi, Japan
[4]Database Center for Life Science, ROIS, Kashiwa-shi, Japan

The BioResource Research Center (BRC) at the Japanese Institute of Physical and Chemical Research (RIKEN) develops and maintains the RIKEN BioResource MetaDatabase. This database integrates several life science datasets to support researchers in making a comprehensive use of RIKEN's research results. For instance, RIKEN BRC collects, preserves and distributes various bioresources for further scientific research: experimental mouse strains, cultured cell lines and genetic materials of human and animal origin. In this article, we present a case study that mainly combines the RIKEN BioResource MetaDataBase [1] and the Bgee, a multi-species gene expression database [2]. Our use case involves finding the Alzheimer disease (AD) related human genes that are highly expressed in the healthy prefrontal cortex, and the RIKEN's genetically modified mice related to these genes. Information on the gene expression is obtained from the Bgee database and the human-mouse orthologs from the Orthologous MAtrix (OMA) database [3]. Orthologs are pairs of genes which have evolved from a single gene in the last common ancestor. The DisGeNET dataset provides relationships between disease and human genes [4].

In summary, we combined the four aforementioned data sources by writing the federated SPARQL query at https://purl.org/data-in-use and illustrated in Fig. 1. This query can be executed in the SPARQL endpoint at https://knowledge.brc.riken.jp/sparql. Examples of retrieved results are shown in Table 1. Among them, the APOE gene is highly relevant for AD research as shown in [5]. Furthermore, we evaluated two query execution scenarios. One scenario considers a SERVICE SPARQL subquery to be executed against the remote Bgee SPARQL endpoint resulting in an average runtime of about 60 seconds and, thus, assuring access to the latest data. The second scenario replaces the SERVICE subquery with a subquery matching triple patterns from the named graph containing Bgee data and stored in the RIKEN BioResource MetaDatabase. As a result, the query runtime is drastically reduced from 60 seconds to around 26 seconds (i.e., 2.3x faster). Therefore, we compared federated versus centralised data access and storage

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences.
*Co-first authors. †Co-last authors.
tarcisio.mendes@sib.swiss (T. Mendes de Farias)

Table 1
Query result example.

Mouse ID	RIKEN's mouse homepage	Ensembl identifier	Gene
RBRC06342	http://www2.brc.riken.jp/lab/animal/detail.php?brc_no=RBRC06342	ENSG00000142192	APP
RBRC03391	http://www2.brc.riken.jp/lab/animal/detail.php?brc_no=RBRC03391	ENSG00000130203	APOE

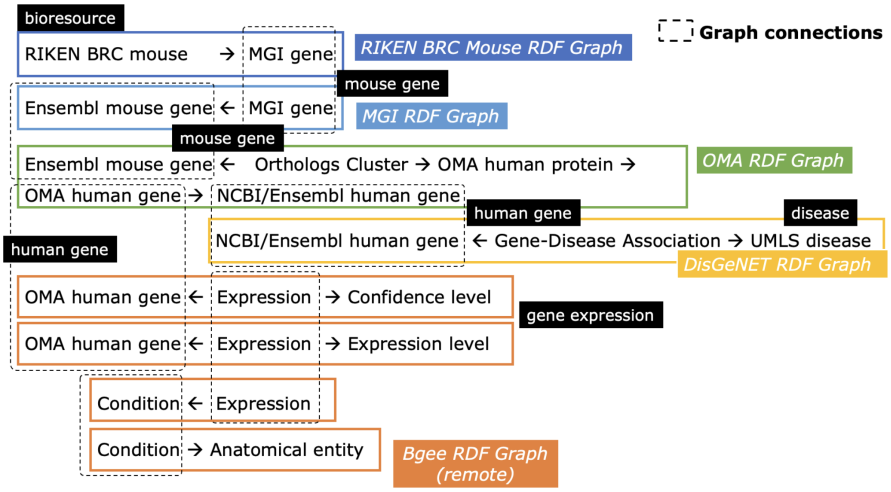

Figure 1: A simplified visualisation of the query graph patterns of our "data in use" example.

approaches for our use case. Although the query execution performance is significantly better in the centralised setup, unfortunately, it does not retrieve the latest results since it currently stores an older version of the Bgee database when compared to the data available via the remote Bgee SPARQL endpoint at https://bgee.org/sparql/. Moreover, the federated setup retrieves 12 more results than the centralised one. To avoid longer runtimes and query timeout, in both scenarios, we used the locally stored OMA and DisGeNET as named graphs in the RIKEN MetaDataBase.

Acknowledgments. Funding from State Secretariat for Education, Research and Innovation (SERI) via ETHZ grant BG 02-072020 and EU Horizon 2020 INODE grant 863410.

References

[1] N. Kobayashi, S. Kume, K. Lenz, H. Masuya, Riken metadatabase: a database platform for health care and life sciences as a microcosm of linked open data cloud, International Journal on Semantic Web and Information Systems (IJSWIS) 14 (2018) 140–164.

[2] 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, et al., The bgee suite: integrated curated expression atlas and comparative transcriptomics in animals, Nucleic Acids Research 49 (2021) D831–D847.

[3] A. M. Altenhoff, C.-M. Train, K. J. Gilbert, I. Mediratta, T. Mendes de Farias, D. Moi, Y. Nevers, H.-S. Radoykova, V. Rossier, A. Warwick Vesztrocy, et al., Oma orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more, Nucleic acids research 49 (2021) D373–D379.

[4] J. Piñero, J. Saüch, F. Sanz, L. I. Furlong, The disgenet cytoscape app: Exploring and visualizing disease genomics data, Computational and structural biotechnology journal 19 (2021) 2960–2967.

[5] K. A. Zalocusky, R. Najm, A. L. Taubes, Y. Hao, S. Y. Yoon, N. Koutsodendris, M. R. Nelson, A. Rao, D. A. Bennett, J. Bant, et al., Neuronal apoe upregulates mhc-i expression to drive selective neurodegeneration in alzheimer's disease, Nature Neuroscience 24 (2021) 786–798.