UNECE High-level Group for the Modernisation of Official Statistics

Business Case for Graph Modeling and Graph Databases support across the Granular Data Lifecycle

This business case was prepared by the Australian Bureau of Statistics and is submitted to the HLG-MOS for their approval.

Type of Activity			
☒	New activity	☐	Extension of existing activity
Proposed Modernisation Group(s) for Activity			
☒	Applying Data Science and Modern Methods	☐	Blue Skies Thinking
☐	Capabilities and Communication	☐	Supporting Standards
☐	*Other:*		

Purpose

The primary purpose of this initiative is to share experiences on how graph modelling and graph databases can help resolve issues associated to the increasing demand of granular data resources by researchers and to the barriers caused by the use of tabular data or relational database solutions when doing data driven policy analysis over multiple administrative and commercial data sources.

Ontologies, RDF triple stores and the SPARQL query language have been adopted by NSOs for metadata management activities (adoption of RDF standards developed by W3C and by the DDI Alliance) and for the publication and consumption of aggregates as *linked data* (ESSNet DIGICOM LOSD). But, so far, there has been limited opportunities in HLG-MOS to share experiences on how their application at earlier stages of the granular data lifecycle can foster:

- The ability to develop, document and maintain complex end to end data pipelines centred on statistical units like Person, Business, Location, through the specification and implementation of declarative mappings of microdata variables to either individual entities or to groups of entities like Households.
- The ability to explain how the data was integrated and disclose its provenance, which statistical disclosure control measures were applied, what measures were taken to make derived granular data products more accessible and of higher value to data analysts.

Areas where the sharing of experience would be extremely valuable include:

- Reasons for selecting a particular graph model (RDF or Property Graph).
- Query performances on large scale graph databases and whether recurring challenges are captured or not in existing benchmarks (LDBC).
- Knowledge Graph Quality Management, in particular for metrics not yet covered in packages designed for the quality assessment of content stored in relational databases or flat files.
- Development environments, including language and tools used to build the graph data model and to construct knowledge graphs but also visualisation and querying tools for helping analysts to learn about the data and manipulate it at a more granular level.

Also of interest, the compatibility and complementarity of graph-based solutions with modern Data Science and Machine Learning (ML) practices such as:

- Federated SQL query engines (Athena)
- Columnar storage formats (Parquet)
- Metadata Management support (Data Catalogs, MLOps) in Data Lakes
- Graph Neural Networks, Large Language Models, Generative AI

Description of the activity and deliverable(s)

The broad objectives of the activity are to:

- Map use cases and analytical scenarios where graph-based approaches are beneficial in particular those associated to Integrated Data Systems (IDS) linking individual level data from multiple agencies.
- Identify barriers which have been stopping NSOs from applying graph-based solutions at various stages of granular data pipelines or to provide graph-enabled analytical capabilities in the secure data access environments provided to external researchers.
- List operations and processes specific to graph-based approaches which have an impact on the quality of the outputs or the cost effectiveness of the solution.
- And finally, increase our insights on the potential benefits of graph-based approaches for future Knowledge-enriched AI services.

For these activities the following steps will have to be executed:

1. Gathering information, best practices, known methods
2. Adapting, integrating and developing into shareable capability building resources

Alternatives considered

NSOs may exchange their experiences on a bilateral basis. NSOs may also engage with one another through other ADSaMM projects or Modernisation Activities.

How does it relate to the HLG-MOS vision and other activities under the Group or HLG-MOS?

This business case aligns with the HLG-MOS vision and values, in particular to accelerating the development of innovative solutions and openly discussing challenges and opportunities.

Proposed start and end dates

Start: January 2024	**End:** December 2024

Graph Modeling and Graph Databases – Background

2023 Maciej Besta et al Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries https://dl.acm.org/doi/pdf/10.1145/3604932 ACM Computing Surveys, June 2023 or https://arxiv.org/pdf/1910.09017.pdf

2023 A Knowledge Representation Approach for Modeling Aggregates: A case study at ISTAT https://www.ital-ia2023.it/submission/61/paper

Mark van der Loo (2022) Topological anonymity in networks https://www.cbs.nl/-/media/_pdf/2022/17/network-anonymity.pdf

2023 The Quest for Schemas in Graph Databases https://amw2023.org/talks/kn/AngelaBonifati-AMW-Keynote-2023.pdf

2022 The World of Graph Databases from An Industry Perspective https://arxiv.org/pdf/2211.13170.pdf

Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A Survey of Functions and Systems. IEEE Transactions on Knowledge and Data Engineering. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10107808

2022 Gu, Z., Corcoglioniti, F., Lanti, D., Mosca, A., Xiao, G., Xiong, J., & Calvanese, D. A systematic overview of data federation systems https://www.semantic-web-journal.net/system/files/swj3074.pdf

2023 Deloitte Responsible Enterprise Decisions with Knowledge-enriched Generative AI https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/risk/deloitte-nl-risk-responsible-enterprise-decisions-with-knowledge-enriched-generative-ai-whitepaper-download.pdf

2023 Reasoning over Financial Scenarios with the Vadalog System https://openproceedings.org/2023/conf/edbt/3-paper-154.pdf

2022 M. Ragab Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads https://dspace.ut.ee/bitstream/handle/10062/88356/ragab_mohamed.pdf https://github.com/DataSystemsGroupUT/SPARKSQLRDFBenchmarking