Blue Brain Knowledge Graph: Leveraging Semantic Web Technologies for Simulation Neuroscience

Cristina E. González-Espinoza[1,*,†], Anna-Kristin Kaufmann[1], Eugenia Oshurko[1], Alejandra Garcia Rojas[1], Sarah Mouffok[1], Konstantinos Platis[1], Jonathan Lurie[1], Jayakrishnan Nair[1], Patrycja Lurie[1], Silvia Jimenez[1], Pierre-Alexandre Fonta[1], Huanxiang Lu[1], Nabil Alibou[1], Mohameth François Sy[1,*,†], Bogdan Roman[1], Samuel Kerrien[1], Henry Markram[1] and Sean L. Hill[1,2,3,*]

[1]*Blue Brain Project, École polytechnique fédérale de Lausanne (EPFL), Biotech Campus, Geneva, Switzerland*
[2]*Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, Canada*
[3]*Department of Psychiatry – Neuroscience and Clinical Translation, University of Toronto, Toronto, Canada*

Abstract

The Blue Brain Project, a Swiss neuroscience research initiative, has pioneered a data-driven approach to digitally building and simulating biologically detailed models of the mouse brain as a complementary approach to understanding the brain alongside experimental, theoretical and clinical neuroscience. One of the key steps of this approach involves acquiring, organizing, and integrating heterogeneous data describing the structural and functional organization of the brain at various levels, ranging from synapses and subcellular components to individual neurons, circuits, and entire brain regions. The data is acquired from many sources including neuroscience experiments, published scientific papers, and brain databases. To address many of the data organization, reuse, sparsity, and publishing challenges that arise alongside this approach, Blue Brain built an RDF-based large-scale knowledge graph bringing together RDFS/OWL ontologies, SHACL schemas, JSON-LD, as well as ontology-, rule-, and graph-based inference to complement classical neuroinformatics tools and methods. In this paper, we present how such a knowledge graph is built and used by the project's domain teams to go beyond high-quality and FAIR metadata cataloging. We describe how the knowledge graph serves a multifaceted role: it addresses the diversity, evolution, and quality assessment of data at the whole brain scale, while concurrently tracking data provenance to facilitate reproducibility and precise attribution. Additionally, it facilitates diverse use cases, including the inference of missing data through knowledge-graph-based methods.

1. Introduction

Blue Brain Project (BBP) has pioneered the new field of *Simulation Neuroscience* using a data-driven and supercomputer-based approach to build and simulate biologically detailed brain tissue models [1]. The approach often involves the steps illustrated in high level terms in Figure 1.

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece
*Corresponding author.
†These authors contributed equally.
crisbeth46@gmail.com (C. E. González-Espinoza); mohameth.sy@epfl.ch (M. F. Sy); sean.hill@epfl.ch (S. L. Hill)
0000-0002-7808-0771 (C. E. González-Espinoza)

[1]https://doi.org/10.1016/j.cell.2015.09.029

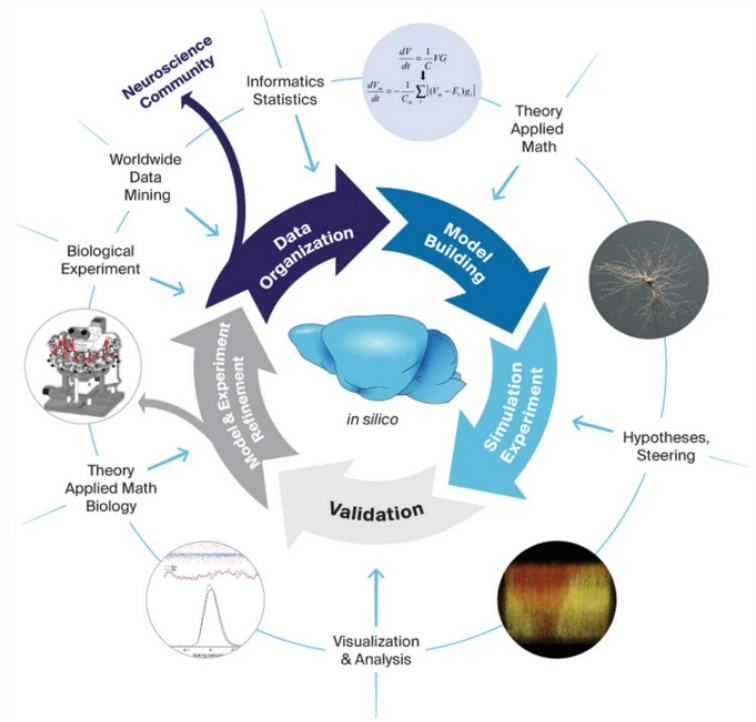

Figure 1: Blue Brain overall Data-driven approach to building and simulating brain tissue models. Source: https://www.epfl.ch/research/domains/bluebrain/research-2/.

In this iterative, data-driven modeling approach, each step informs and refines the subsequent one, giving rise to a set of challenges concerning the storage, accessibility, and reusability of data and models. These challenges manifest as data is acquired, generated, or published both internally for users and to a wider scientific audience. For instance, during the data organization step, the discovery, acquisition, preparation, and release of multi-scale, multi-modal, and heterogeneous data are essential for enabling access and facilitating reuse in model development. Furthermore, tracking data and models provenance is key during the model validation step to select validation data different from the one used during model building but also during model publication to support reproducibility and for contribution attribution and quality assessment.

2. Knowledge graph as an approach for organizing neuroscience data

The heterogeneity challenge comes often from data of different sizes, formats, generation contexts and sources. For example, to build single neuron model at a given brain region location (e.g somatosensory cortex, hippocampus, thalamus), a neuroscientist modeler would often need to get neuron morphologies (i.e. reconstructed neuron 3d shape) and electrophysiological recordings (i.e measured electrical behaviour) from which to extract features. While many

neuroscience databases collect and enable the modeler to search and download neuron morphologies (e.g. NeuroMorpho.org, Mouselight) or neuron electrophysiological recordings (e.g. Allen Cell type DB), they greatly vary in term of (meta)data formats and on accounting on the data generation contexts.

These challenges are not specific to simulation neuroscience and are summarized in the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for scientific data management[2]. Addressing these challenges is a clear target and use case for Semantic Web technologies[1]. BBP built a knowledge graph, the Blue Brain Nexus [2], by leveraging semantic web technologies and deploying it as an ecosystem or platform, to support building and simulating brain tissue models[3]. Nexus provides a complement to classical neuroinformatics for organizing neuroscience data. Furthermore, Nexus is domain agnostic and can be used in any data-driven field as a knowledge graph technology stack.

2.1. Building the knowledge graph from different data sources

At Blue Brain, building a knowledge graph from difference sources can be summarized in three main steps: i) define, in the form of W3C SHACL [3] and ontologies, the schemas and formats of the targeted neuroscience entities' types, their metadata referring to the key scientific, technical activities, protocols, and agents involved in their generation; ii) define simple declarative JSON-based transformation or mapping rules to map source data to targeted schemas; and iii) apply the mappings to the data from a given source and register the results in the knowledge graph. These mappings[4] are used with Nexus Forge[5], a Python framework for building knowledge graphs.

In order to exemplify the complexity of the data integration process, figure 2 shows two schematic diagrams of two very different generation contexts of the same type of neuroscience data: a neuron morphology.

2.2. Knowledge graph schema

In Blue Brain, many SHACL shapes and ontologies [6] have been developed as the knowledge graph schema, extending existing standards such as schema.org and W3C PROV-O. The shapes and ontologies cover entities from the subcellular level to the whole brain such as neuron morphologies, electrophysiological recordings, ion channel recordings, parameters from literature, brain atlases, cell composition of the brain, brain regions, cell types, species, etc. The W3C RDF format is leveraged, specifically its developer-friendly JSON-LD serialization, which eases federated access and discoverability of distributed neuroscience (meta)data over the web.

[2]https://doi.org/10.1038/sdata.2016.18
[3]https://www.w3.org/TR/shacl/
[4]https://github.com/BlueBrain/bbp-ontologies/tree/master/mappings
[5]https://nexus-forge.readthedocs.io/en/latest/interaction.html#mapping
[6]https://github.com/BlueBrain/bbp-ontologies

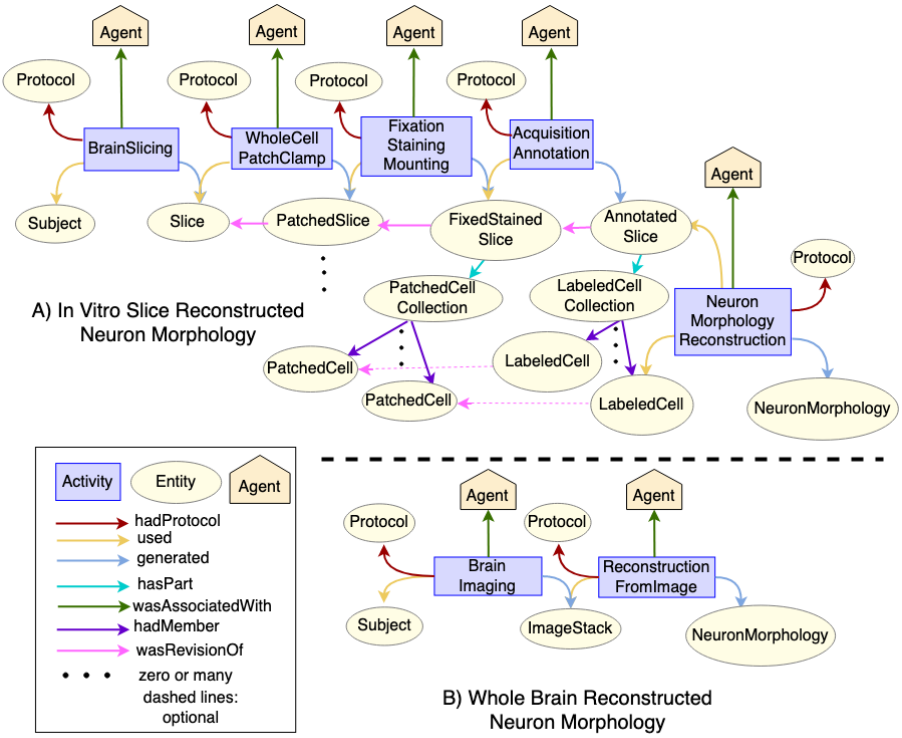

Figure 2: *Neuron Morphology*, the same neuroscience data type but different data sources and generation contexts: A) In Vitro Slice Reconstructed Neuron Morphology from Allen Cell type DB, Neuromorpho.org; B) Whole Brain Reconstructed Neuron Morphology from Mouselight

2.3. Publishing neuroscience data on the web

The organized, linked and curated data can be shared both internally and externally as web portals dynamically built from SPARQL and ElasticSearch queries sourced from the knowledge graph. The same data is also published as programmatically accessible knowledge graphs. An example is the Thalamoreticular Microcircuitry web portal [7] allowing users to browse, visualize, query and download the experimental data (e.g. 3D neuron morphologies, electrophysiological recordings, and interactive visualizations) used to build digital reconstructions (e.g. single cell model and microcircuit reconstruction) as well as network simulations.

2.4. Inference as a tool for neuroscience data generalization

One of the main challenges in simulation neuroscience is the sparsity of the data. For instance, certain brain regions have received very little attention by experimentalists and therefore very few neuronal morphologies and electrophysiological recordings appear in the literature for these areas. In this situation, scientists may want to find and adapt the same types of data but from a different but "similar" brain regions, cell types or species (e.g. borrow or adapt

[7] https://bbp.epfl.ch/portals/thalamoreticular

rat data for building mouse models). These type of *data generalizations* can be expressed in the form of knowledge graph-based inference rules. Three main strategies are followed: 1) classical ontology-based generalization (e.g. a parent brain region can be considered similar to its descendants); 2) knowledge graph embeddings obtained using metadata and graph structures (eg. embeddings of neuron morphologies are built from their neighbours in the graph using techniques such as RDF2VEC [4]); and 3) similarity embeddings generated from entity features (eg. an embedding vector is built for each neuron morphology by vectorizing its 3D shape using topological techniques [5]).

3. Conclusion

In this work, a complex use-case for semantic web technologies has been presented in the context of the emerging field of simulation neuroscience. This particular domain poses a formidable challenge due to the heterogeneous nature of data sources, the data sparsity, and the dynamic nature of terminology and conceptual models.

Acknowledgments

This study was supported by funding to the Blue Brain Project, a research center of the École polytechnique fédérale de Lausanne, from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology. Funding has been provided in part from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2).

References

[1] P. Ristoski, H. Paulheim, Semantic Web in data mining and knowledge discovery: A comprehensive survey, Journal of Web Semantics 36 (2016) 1–22. URL: http://www.sciencedirect.com/science/article/pii/S1570826816000020. doi:10.1016/j.websem.2016.01.001.

[2] M. F. Sy, B. Roman, S. Kerrien, M. D. M., H. Genet, W. Wajerowicz, M. Dupont, I. Lavriushev, J. Machon, K. Pirman, D. Neela Mana, N. Stafeeva, A.-K. Kaufmann, H. Lu, L. Jonathan, P.-A. Fonta, A. G. R. Martinez, A. D. Ulbrich, C. Lindqvist, S. Jimenez, D. Rotenberg, H. Markram, S. L. Hill, Blue brain nexus: An open, secure, scalable system for knowledge graph management and data-driven science, Semantic Web 14 (2022) 697–727. URL: http://doi.acm.org/10.3233/SW-222974. doi:10.3233/SW-222974.

[3] F. Schürmann, J.-D. Courcol, S. Ramaswamy, Computational Concepts for Reconstructing and Simulating Brain Tissue, Springer International Publishing, Cham, 2022, pp. 237–259. URL: https://doi.org/10.1007/978-3-030-89439-9_10. doi:10.1007/978-3-030-89439-9_10.

[4] P. Ristoski, H. Paulheim, RDF2vec: RDF Graph Embeddings for Data Mining, in: International Semantic Web Conference, 2016. doi:10.1007/978-3-319-46523-4_30.

[5] L. Kanari, P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess, H. Markram, A topological representation of branching neuronal morphologies, Neuroinformatics 16 (2018). doi:10.1007/s12021-017-9341-1.