A Graph-based Approach for Representing, Integrating, and Analysing Neuroscience Data: The Case of the Murine Basal Ganglia

Maren Parnas Gulnes[1], Ahmet Soylu[2], and Dumitru Roman*[1]

[1]SINTEF AS, Norway
[2]OsloMet - Oslo Metropolitan University, Norway

Abstract

Purpose: Neuroscience data is spread across a variety of sources, typically provisioned through ad-hoc and non-standard approaches and formats, and often has no connection to the related data sources. These make it difficult for researchers to understand, integrate, and reuse brain-related data. The aim of this study is to show that a graph-based approach offers an effective mean for representing, analysing, and accessing brain-related data, which is highly interconnected, evolving over time, and often needed in combination.
Approach: We present an approach for organising brain-related data in a graph model. The approach is exemplified in the case of a unique data set of quantitative neuroanatomical data about the murine basal ganglia — a group of nuclei in the brain essential for processing information related to movement. Specifically, the murine basal ganglia data set is modelled as a graph, integrated with relevant data from third-party repositories, published through a Web-based user interface and API, analysed from exploratory and confirmatory perspectives using popular graph algorithms to extract new insights.
Findings: The evaluation of the graph model and the results of the graph data analysis and usability study of the user interface suggest that graph-based data management in the neuroscience domain is a promising approach, since it enables integration of various disparate data sources, and improves understanding and usability of data.
Originality: The study provides a practical and generic approach for representing, integrating, analysing, and provisioning brain-related data, and a set of software tools to support the proposed approach.

Keywords: Graph databases, neuroscience, brain-related data, murine basal ganglia, data integration, data analytics, and data visualisation.

*Corresponding author: dumitru.roman@sintef.no.

DOI: https://doi.org/10.1108/DTA-12-2020-0303

1 Introduction

The brain is the organ humans rely on the most but understand the least. In order to understand the brain's structure and function, researchers need data. In this respect, neuroscience data, primarily representing the features of the brain and information related to brain-related research, has increased significantly over the past decade due to the advances in technology (Fan & Markram (2019)). The data that exists about the brain already is in large quanta, complex, and spread across repositories in multiple formats. As an example of this complexity, brain-related data can represent a part of the human brain's 86 billion neurons, and for each neuron, any of the approximately 7000 connections (synapses) (Herculano-Houzel (2009), Drachman (2005)). The existing neuroscience data, however, is spread across a variety of sources, typically provisioned through ad-hoc and non-standard approaches and formats, and often has no connection to the related data sources (Bassett et al. (2018)). These primarily hinder the reuse, integration, and sharing of data (Teeters et al. (2008)) and it becomes increasingly difficult for researchers to combine and use relevant data. Therefore, there is a need to examine how neuroscience data can be modelled and stored to facilitate integration and reuse. Existing research on managing brain-related information mostly works towards the standardisation of metadata, aiming to make it easier for researchers to find and reuse data (Amunts et al. (2016), Gardner et al. (2012), Sivagnanam et al. (2013)). This research stream, however, mainly focuses on metadata management for data sets and little on managing the actual data.

In this respect, graph data models and databases provide performance, flexibility, and agility (Fernandes & Bernardino (2018)) and open up the possibility of using well-established graph analytics solutions (Needham & Hodler (2019)); however, there is little research on graph-based data representation as a mechanism for integration, analysis, and reuse of neuroscience data. Therefore, in this article, we address the following research questions:

1. Can graph-based representation of brain-related data facilitate the integration of data from a variety of neuroscience data sets?

2. Can a graph model provide a better understanding of the data in a brain-related data set?

3. To what extent can a graph-based approach to neuroscience data management improve the usability of the data?

To this end, in this article, we show that a graph-based approach offers an effective mean for representing, analysing, and accessing brain-related data by presenting an approach for organising brain-related data in a graph model enabling integration of various disparate data sources, and improving the understanding and usability of data. The approach is exemplified in the case of a unique data set of quantitative neuroanatomical data about the murine basal ganglia (Bjerke et al. (2020)) — a group of nuclei in the brain essential for processing information related to movement. Specifically, the murine basal ganglia data set is modelled as a graph, integrated with relevant data from third-party repositories, published through a Web-based user interface and API, and analysed from exploratory and confirmatory perspectives using popular graph algorithms to extract new insights. Through exploratory and confirmatory analysis, we managed to find interesting findings and answer specific questions. The evaluation of the graph model and the results of the graph data analysis and usability study of the user interface suggest that graph-based data management in the neuroscience domain is a promising approach. The study presented in this article provides a practical and generic approach for representing, integrating, analysing, and provisioning brain-related data, and a set of software tools to support the proposed approach.

The rest of the article is structured as follows. In Section 2 provides the background information, while Section 3 presents the related work. Section 4 describes the data sets and Section 5 presents the design and implementation of the proposed solution. Section 6 presents the evaluation, while finally, Section 7 concludes the article.

2 Background

In this section, we provide some background information on neuroscience and graph-based data representation in order to facilitate comprehensibility of rest of the article for researchers with different backgrounds (e.g., neuroscience and computer science).

2.1 Neuroscience

The brain is a large and complex organ that, together with the spinal cord, constitutes the central nervous system (CNS) (Kandel et al. (2000)). Neuroscience typically divides the brain into different parts based on each region's functional, connectional, or structural properties. The exact division varies across the literature, but Kandel et al. (2000) specify six main parts. These can be grouped into the three parts presented, namely cerebrum, brainstem, and cerebellum. Cerebrum consists of the cerebral cortex and subcortical nuclei. The cerebral cortex is the outer layer of the cerebrum and is responsible for most human cognitive abilities. The subcortical areas lie, as the name suggests, beneath the cortex. It consists of three compounds where one of them is the basal ganglia (Kandel et al. (2000)). The brain includes numerous different cell types, broadly categorised as glial cells and neurons (Campbell et al. (2011)). Neurons are the cells that process and transport information throughout the CNS. They communicate through connections called synapses. The brain glial cells are non-neural, meaning they do not transfer signals directly. Instead, the glia cells provide support and regulate the functioning of the neurons (Campbell et al. (2011)). Much of brain-related research investigates the cells in the brain. Many disease studies use rats or mice for their research, since rodents have a shorter lifespan and researchers can observe them in controlled environments.

There are differences in the division, region naming, and which parts of the brain a defined region contain (Swanson (2000)). When a neuroscientist makes an observation, it is vital to communicate the observation's location in the brain (Bjerke et al. (2018)). In science, a nomenclature defines a system for naming within a specific area (Merriam-Webster.com dictionary (n.d.)). In neuroscience, a brain region nomenclature is a framework for naming and defining the areas of the brain. When studying the brain, such nomenclatures help researchers precisely define which region or part of the brain the data reference (Bowden et al. (2012)). Neuroscience researchers utilise brain atlases for matching the location of their findings. In this study, we refer to the term "brain atlas" as the more narrow description of atlases used for reference, also called reference atlases. A reference brain atlas is a map of the brain for a specific species, containing images of the brain and borders between regions in the context of those images (Bjaalie (2002)). In relevance to anatomical naming, reference atlases employ a specified nomenclature (Bjaalie (2002)). The nomenclatures of the most renowned brain atlases at a given time are what researchers usually choose as nomenclature in a study or research experiment (Bota & Swanson (2010)). For example, when measuring cell-counts in a region, researchers can report which atlas nomenclature they have used to specify the given region. That atlas nomenclature is then the nomenclature used in that experiment or research. This reporting is essential for other researchers to obtain the correct location of the research observations. Another area of anatomical naming considers cell types. Neuroscience research is often not concerned with counting or observing all neurons, but rather specific neurons, such as neurons which express particular neurotransmitters (Shepherd et al. (2019)). Researchers can name the neurons based on what they express, where they exist in the brain, or their structure, based on the research focus (Hamilton et al. (2012), Petilla Interneuron Nomenclature Group (PING) (2008), Shepherd et al. (2019)). The many ways researchers can describe a cell type cause a lack of consensus on the criteria for defining neuron types (Hamilton et al. (2012)). For clarity, researchers should explicitly report what defines a specific cell type in their research (Shepherd et al. (2019)).

The basal ganglia are not a concrete part of the brain but a collective term for a group of nuclei. In humans and other mammals, the basal ganglia are significantly involved with movement and, to some degree, emotions, and memories (Gerfen & Bolam (2016), Middleton & Strick (2000)). Much of the basal ganglia's clinical significance is related to movement disorders like Huntington's disease and Parkinson's disease (Obeso et al. (2008), Bunner & Rebec (2016)). Basal ganglia studies are often related to specific diseases, producing a predominance of data about brain regions and cell types relevant to the disease.

2.2 Graph-based data representation

Many real-world scenarios are naturally structured as graphs, such as social networks and neuron connectivity. Graph-based data representation provides a way to represent such real-world structures directly. Graph-based data representation entails all representations of data that utilise a graph model. There are various types of graphs. Discrete mathematics defines a graph, or a simple graph, as a set of vertices (nodes) and edges (relationships). Nodes connect through edges, and all edges in a graph go between two nodes in the node-set (Cormen et al. (2009), Diestel (2017)).

Graph databases are utilised to use graph models. Robinson et al. (2013) define a graph database as a database management system with Create, Read, Update, and Delete (CRUD) methods that expose a graph data model. When defining graph databases, there is a separation between native and non-native implementations (Needham & Hodler (2019), Fernandes & Bernardino (2018)). A native graph database as a graph database that has a graph data model in the underlying storage. It processes the data using index-free adjacency, meaning that the connected database entries (nodes) point to each other's physical location (Robinson et al. (2013)). A relational data model can also be viewed as a graph, but with limitations. ER diagrams, commonly used to model and presents relational databases, are graphs where the tables represent nodes, and the foreign-keys define named relationships. The graph databases' ease of changing schemas provides flexibility; graph databases do not have strict predefined schemas that all nodes of a particular type need to follow. Instead, one can define what needs to be there as the database and application evolve, representing the domain model. The non-strict schemas also supply the agility advantage, allowing the database to change with the domain requirements (Fernandes & Bernardino (2018)). In this study, we focus on NoSQL databases (Han et al. (2011), Cattell (2011)). NoSQL databases perform fast read and write procedures, support large data sets, and deal well with dynamic data, both changes in the schema and the data size (Robinson et al. (2013), Han et al. (2011), Cattell (2011), Hecht & Jablonski (2011)).

A graph database exposes a graph data model. Angles & Gutierrez (2008) define a graph database model as a model where the data structure (schema) is modelled as a graph and where the data manipulation uses graph-based operations. There are many different graph data models, but the two most common are the property graph model and the RDF graph model. A property graph model is a graph model with nodes and relationships where both the nodes and the relationships can have properties. The model categorises the nodes with one or more labels, and the relationships are named and directed (Robinson et al. (2013)). The RDF is a standard model, developed under the World Wide Web Consortium (W3C), enabling the encoding, exchange, and reuse of structured metadata Miller (1998). The goal was to make a framework for all the World Wide Web resources to improve programmatic discovery and access to these sources (Miller (1998), Pérez et al. (2009)).

Graph analytics includes all approaches to analyse graph-based data (Needham & Hodler (2019)). Scarselli et al. proposed a graph neural network (GNN) model that utilised existing neural network methods on data represented in a graph model (Scarselli et al. (2008)). Graph neural networks (GNNs) have gained some use over the past decade (Zhou et al. (2018)). There are many scenarios for using GNNs to predict and classify graph data models. Some are related to traditional machine-learning tasks, such as models for text and image classification. Other scenarios are more specific to data naturally structured as graphs, such as disease classification, protein interface prediction, and knowledge graph completion and alignment (Zhou et al. (2018)). Traditional graph algorithms include path finding and search, community detection, centrality, and similarity. Path finding and search algorithms are concerned with graph search by traversing the graph (Needham & Hodler (2019)). Community detection algorithms discover communities in a graph (Needham & Hodler (2019)). Centrality algorithms measure which nodes are the most influential and have an extensive impact on the graph (Newman (2018)). Finally, similarity algorithms measure the similarity of nodes by comparing node pairs (Newman (2018)).

3 Related work

Researchers generally separates data into multiple types (e.g., clinical and genetic) and modalities (e.g., species and diseases) and process data at different levels National Academies of Sciences, Engineering, and Medicine (2020). From the data processing point of view, the data could be categorised into three levels: (1) raw data, (2) derived data, and (3) metadata. Raw data entails direct research measurements and can be neuroimages, electrode recordings, or other direct measurements. Researchers analyse the raw data to provide insights and this process yields the derived data. Examples of derived data are quantitations (objects of interest counts), distributions, or morphologies (an object's physical structure). Metadata defines the characteristics of this data, being the "data about data". Metadata covers all the information related to an experiment and can include data about the methodology, specimens, and specific chemical solutions of the research. There are several initiatives for sharing data in order to advance brain research. Many of the initiatives are complementary and attempt to build on each other's data.

For example, there exist research on creating common frameworks for neural data. Hamilton et al. (2012) proposed an ontological approach for describing neurons and their relationships. Due to the numerous ways research can identify neurons, it is unlikely that a standard naming format for the data can exist. Consequently, research and data initiatives have created guidelines on how to handle the data, with the central notion being the data must be made available and machine-readable (Akil et al. (2011), Ascoli (2012)). Therefore, in the followings, we review initiatives focusing on providing data and graph-based approaches in neuroscience. None of the approaches reviewed integrate disparate research data while maintaining the metadata, which is the aim of this study.

3.1 Neuroscience data intiatives

We first consider repositories of data sets, which entails initiatives collecting and providing research data from multiple sources, including publications and data sets. The goal is to make neuroscience research more available to deal with the data quality challenges. Some of these are general-purpose, and others are specialised for a research area or data set (Amunts et al. (2016), Sicilia et al. (2017), Gardner et al. (2012), *KnowledgeSpace* (n.d.)). Another essential brain-related data initiative type is brain atlases. We particularly focus on reference brain atlases, which are maps of the brain, including defined brain region borders. Researchers use these atlases as reference tools to answer questions about location in the brain (Bjaalie (2002)). In comparison to the repositories of data sets, atlases come from one data source. Although brain atlases do not integrate research data directly, they are essential for neuroscience data integration as they provide location references for research and standardisation of these locations (Bjerke et al. (2018)). The final initiative type that we consider is neuroscience databases. A neuroscience database is broadly a database consisting of brain-related data. These initiatives integrate data from one or many sources, such as research papers or other databases, into a common database (at any or all data levels). The murine basal ganglia database is an example of such an initiative (Bjerke et al. (2019)).

3.2 Graph-based approaches

Graph-based data representation in neuroscience has primarily focused on knowledge graphs for organising research and networks for the brain's neural connections. We can consider two main approaches to graph-based data representation in neuroscience as two sides of a scale. On the one side, we have the knowledge graph approaches that utilise graph models for managing research metadata to integrate multiple data sets. On the other side, we have approaches that integrate research data into a complete model and remove all metadata references. The first type of brain-related initiatives include EBRAINS and KnowledgeSpace utilising knowledge graphs to enrich data and improve search engines that retrieve research papers and data sets (*EBRAINS* (n.d.), *KnowledgeSpace* (n.d.)). Both initiatives provide powerful graph models that simplify data discovery; however, they do not change or connect the data in the papers, data sets, and models contained in the knowledge base. Another direction of computational neuroscience that utilises graph principles is the study of neural connections in the brain, called connectomics. Connectomics is an extensive research field, including numerous research papers and large initiatives, such as the five-year Human Connectome Project (Van Essen et al. (2013)). Connectomics presents an example of data that is naturally structured as a graph and that can benefit from graph-based data representation; however, it has no direct reference to the employed techniques.

4 Data sets

In this section, we first describe the murine basal ganglia data set, the main data set used in this study. Then we describe our efforts for finding related data sets that could be integrated with the murine basal ganglia data set and the results.

4.1 The murine basal ganglia data set

The murine basal ganglia data set, created by Bjerke et al. (2020), consists of quantitative neuroanatomical data about the healthy rat and mouse basal ganglia, collected from more than 200 research papers and data repositories. The data set contains three distinct information types: quantitations (counts), distributions, and cell morphologies. The counts and distributions regard either entire cells or specific parts of the cell, while the morphologies describe the cell's physical structure. The data set is publicly available through EBRAINS as an Access database and a set of CSV-files (Bjerke et al. (2019)). The data set's primary usage is for researchers to find and compare neuroanatomical information about the basal ganglia brain regions. In addition to the data set, Bjerke et al. (2020) published a paper describing the data set development process and their findings.

The data set contains metadata and derived data. The murine basal ganglia data set contains metadata about the experiments, analyses, and specimens, connected to experimental results and derived data, representing cell counts and cell structures in specific regions. Moreover, it contains general cell types and brain regions used to reference the derived research results. All the data in the murine basal ganglia data set are in a tabular format. The brain regions in the data set come from two nomenclatures, provided by the Waxholm Space (WHS) rat brain atlas for the rat species and the Allen Mouse Brain Atlas (AMBA) for the mouse species. The data set does not contain raw data or externally referenced files with data.

Figure 1 illustrates the murine basal ganglia database directly represented as a conceptual graph, where nodes represent the database tables, and edges represent the foreign keys. It depicts the data structure, including the connectivity information. The nodes are marked with one of four colours. These colours represent four node-categories we derived from investigating the data in discussion with a neuroscience expert. The following list presents these categories with associated colours:

- *Experiment data* (purple): The nodes representing experiments and the related experiment data.

- *Sources of information* (green): The nodes representing external sources of information. This category includes the sources (journals) that published the experiments and the nomenclatures used to define the brain regions.

- *Specimen data* (yellow): The nodes representing the experiment's specimens and the properties of these.

- *Neuroanatomical data* (orange): The nodes representing neuroanatomical data about brain regions and cells with classifications and areas.

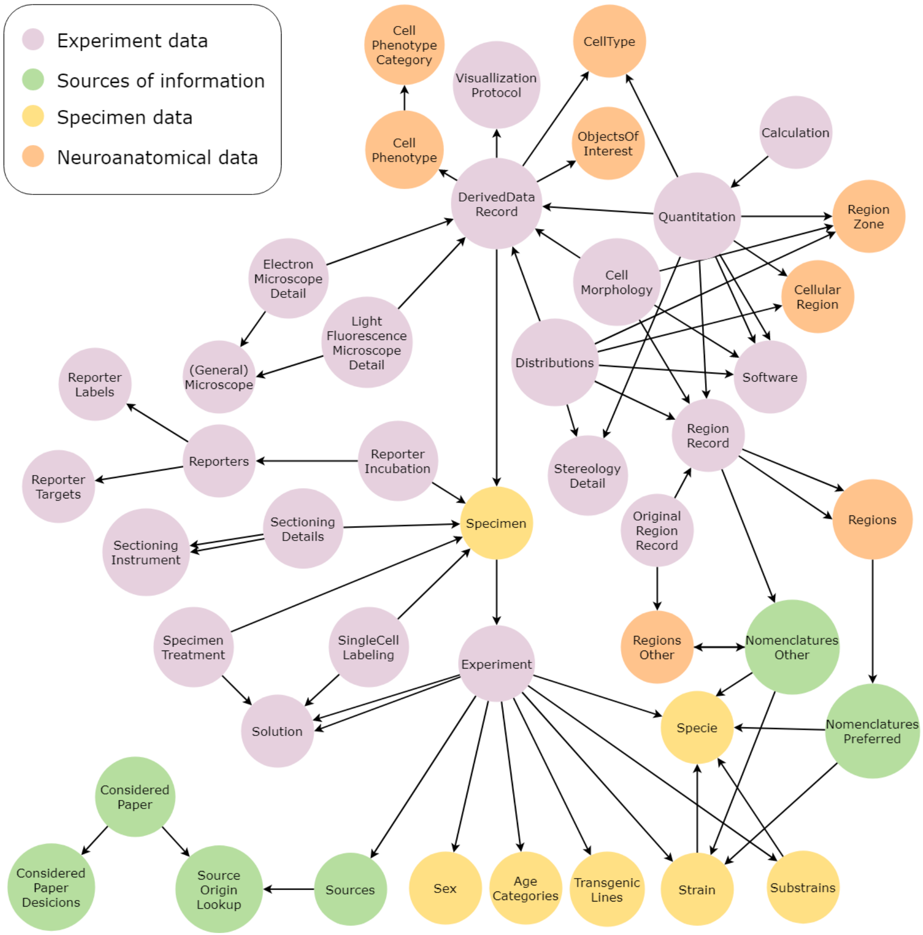


Figure 1: The structure of the original murine basal ganglia database, presented as a graph.

4.2 Related data sets

We evaluated of a set of related initiatives that publicly provide neuroscience data in order to identify data sets that overlap with the murine basal ganglia data set for the integration purposes. This section does not include a complete overview of all available neuroscience data initiatives and does not perform an in-depth examination of each source. Instead, it provides an examination of what data we can obtain and integrate from such initiatives. The overall methodology to investigate the initiatives was to visit the data source and look for available data sets and programmatic data access. For the initiatives that provided data programmatically, we searched with the term “basal ganglia” and some specified basal ganglia related regions. Where we managed to obtain relevant data, we consulted an expert to evaluate if the data was related to the basal ganglia. Further, we evaluated if the data was overlapping with the murine basal ganglia data set. If these criteria were met, the data could be integrated into the murine basal ganglia graph model. In summary, we analysed each initiative against the following criteria: (i) serves data programmatically, (ii) has data related to the basal ganglia, (iii) and provides data that could be connected to murine basal ganglia.

The intiatives considered were: (1) EBRAINS[1] (repository, multiple species), (2) Neuroscience Information Framework[2] (repository, multiple species), (3) Zenodo[3] (repository, multiple species), (4) Knowledge-Space[4] (repository, multiple species) (5) Waxholm Space (WHS) rat brain atlas[5] (atlas, rat), (6) Allen Institute for Brain Science[6] (atlas and repository; human and mouse), (7) The Blue Brain Cell Atlas (BBCA)[7](data set, multiple species), (8) Brain Architecture Management System (BAMS)[8](data set, rat), (9) NeuroMorpho.Org[9] (Data set, multiple species), (10) and InterLex through SciCrunch[10] (repository data set, multiple species).

Figure 2 summarises our investigation, presenting data sources that we can collect data from. Brain Architecture Management System (BAMS) contains basal ganglia related information extendable with the murine basal ganglia data set; however, it does not provide the data programmatically. BAMS Website presentes the data in tables, that could be accessed programmatically to extract the data. InterLex through SciCrunch provides an API[11] where one can get descriptions based on InterLex's ontological identifiers. In the murine basal ganglia data set, many of the cell types have ontological identifiers recorded. These identifiers could be used to connect the information to the cell types in the database. NeuroMorpho.Org provides an API[12] where researchers can find a neuron by id or name. With the inspiration of looking for identifiers in the murine basal ganglia data set, we observed that the nodes with cell morphologies (the structure of the cell) also had identifiers for the neurons mapping to NeuroMorpho.Org. It is important to note that this is not an in-depth analysis. Some of the initiatives that we found unsuited for integration might fulfil all the criteria, but not in a way we managed to observe at the time of the study.

[1] `https://ebrains.eu`

[2] `https://neuinfo.org`

[3] `https://zenodo.org`

[4] `https://knowledge-space.org`

[5] `https://www.nitrc.org/projects/whs-sd-atlas`

[6] `https://alleninstitute.org/what-we-do/brain-science`

[7] `https://portal.bluebrain.epfl.ch/resources/models/cell-atlas`

[8] `https://bams1.org`

[9] `http://neuromorpho.org`

[10] `https://scicrunch.org/scicrunch/interlex/dashboard`

[11] `https://scicrunch.org/browse/api-docs/index.html?url=https://scicrunch.org/swagger-docs/swagger.json`

[12] `http://neuromorpho.org/apiReference.html`

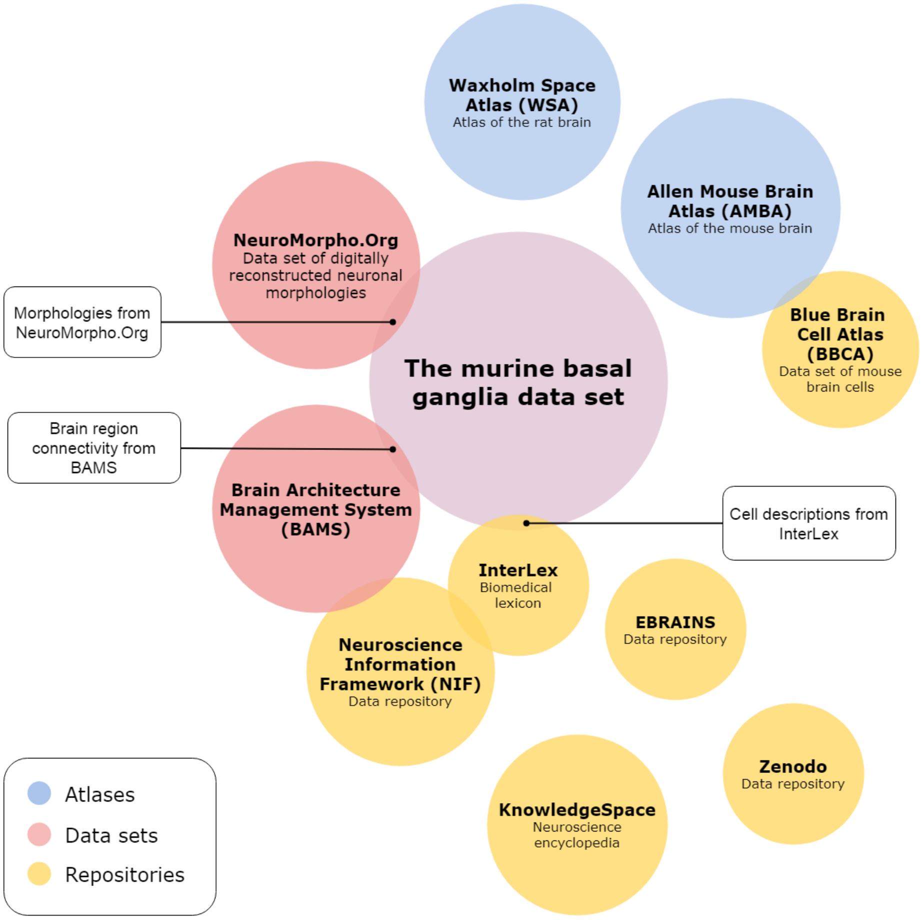

Waxholm Space Atlas (WSA)
Atlas of the rat brain
Allen Mouse Brain Atlas (AMBA)
Atlas of the mouse brain
NeuroMorpho.Org
Data set of digitally reconstructed neuronal morphologies
Morphologies from NeuroMorpho.Org
The murine basal ganglia data set
Blue Brain Cell Atlas (BBCA)
Data set of mouse brain cells
Brain region connectivity from BAMS
Brain Architecture Management System (BAMS)
Cell descriptions from InterLex
InterLex
Biomedical lexicon
EBRAINS
Data repository
Neuroscience Information Framework (NIF)
Data repository
Zenodo
Data repository
KnowledgeSpace
Neuroscience encyclopedia
● Atlases
● Data sets
● Repositories

Figure 2: Overview and result of initiatives investigated for data overlap with the murine basal ganglia data set.

5 Design and implementation

In this section, we describe the proposed solution, including design and implementation, for graph-based data modelling and integration of neuroscience data. Software components brought together for data analytics and a Web interface developed for end-user access to data is described in this section as well.

5.1 Architecture

Figure 3 depicts a high level overview of the architecture with different components. We designed and implemented a graph model for the murine basal ganglia data set, chose a Graph Database Management System (GDBMS), and migrated the data from the relational database into the graph database, Neo4j.

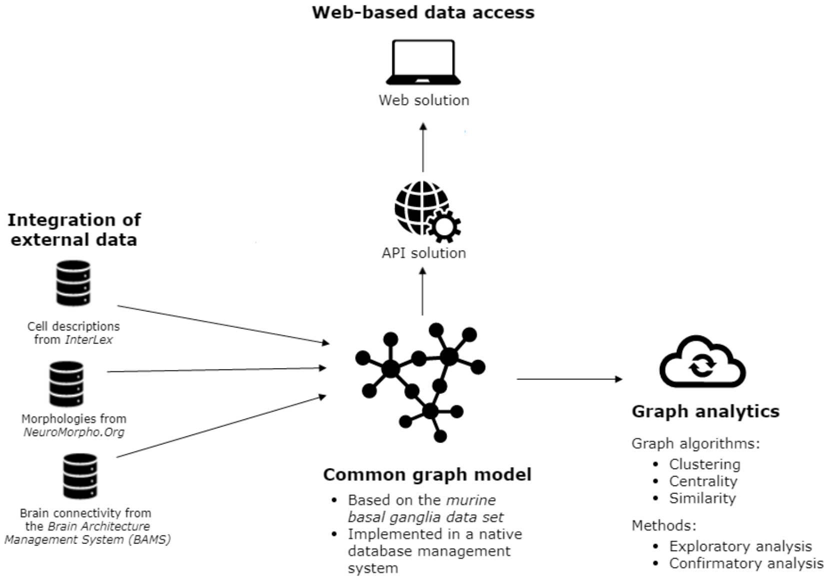

Web-based data access
Web solution
API solution
Integration of external data
Cell descriptions from *InterLex*
Morphologies from *NeuroMorpho.Org*
Brain connectivity from the *Brain Architecture Management System (BAMS)*
Common graph model
• Based on the *murine basal ganglia data set*
• Implemented in a native database management system
Graph analytics
Graph algorithms:
• Clustering
• Centrality
• Similarity
Methods:
• Exploratory analysis
• Confirmatory analysis

Figure 3: High-level architecture of the proposed solution.

Further, we designed and implemented the integration of data from related neuroscience data sources and the technical solution for graph-based data analysis. To provide Web-based access to the graph data, we designed and implemented a Web application and API. The following list presents the main components in Figure 3:

1. *The common graph model*: It is based on a native database management system.

2. *Integration of external data*: This includes external data sources as identified earlier.

3. *Graph analytics*: Tools and approaches used to analyse the graph data based on the techniques described earlier.

4. *Web-based data access*: A Web-based access interface to the data based on the graph model.

5.2 Graph-based data modelling

We designed a graph model for the murine basal ganglia data set. Figure 4 presents the high-level design of the graph model based on the murine basal

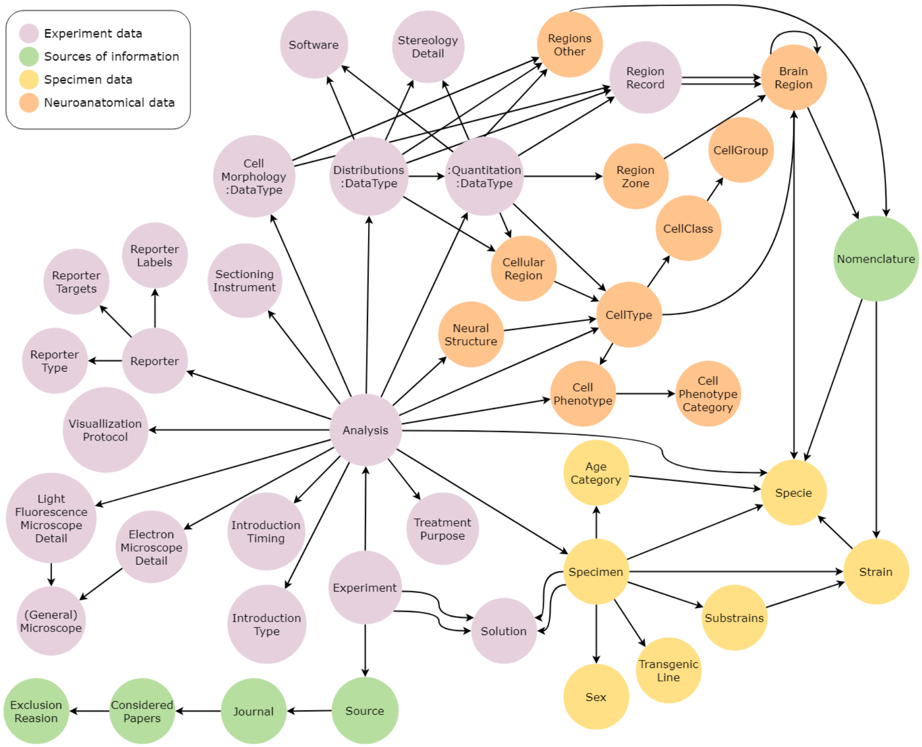


Figure 4: The high-level design of the graph model of the murine basal ganglia data set.

ganglia data set. We involved researchers from the Faculty of Medicine at the University of Oslo, while designing the data model.

In addition to specific design choices in consultation with the experts such as new relationships and node labels, we followed a general approach[13] for converting a relational database model into a graph model: (i) a table becomes a node label; (ii) each row in the table becomes a node of that label; (iii) each column of the row becomes a property of the node; (iv) foreign keys become relationships; (v) and join-tables become relationships with properties. The resulting graph model is a directed multigraph, since the relationships have direction and some node pairs have multiple relationships. It also contains a self-loop on the brain region node type. Further, the graph is connected, as there is a path between all the nodes in the graph. We decided to model the graph after what was appropriate for the domain model and requirements and chose to adjust analysis methods accordingly.

[13] https://neo4j.com/developer/relational-to-graph-modeling

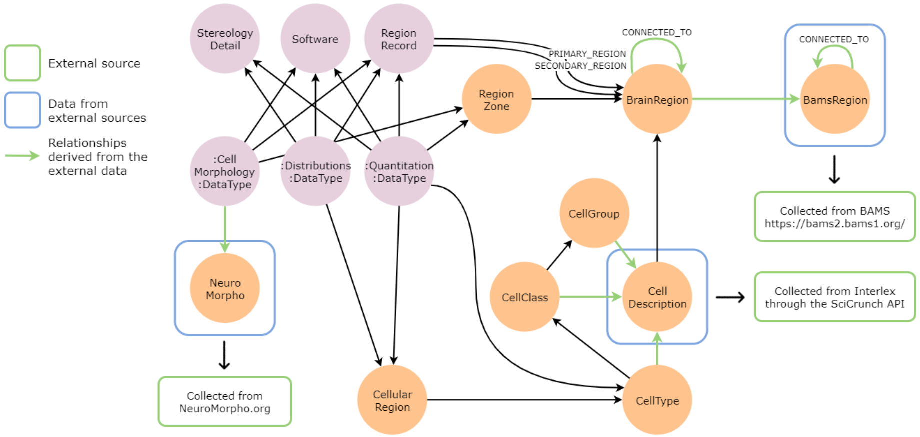


Figure 5: Integration of data from external sources.

5.3 Data onboarding and integration

We onboarded data from the relational database to a graph database. We employed Neo4j as it implements native graph storage, provides integration with many programming languages, has fast create-retrieve-update-delete procedures, is popular and well documented, and can apply graph algorithms on the data. Extract-Load-Trasnform (ETL) tools help extract data from a source, transform it to fit the destination database's schema, and load it into the destination database (Özsu & Valduriez (2019)). Neo4j provides such as tool, Neo4j ELT[14], that automatically maps a relational database to a graph database. However, our graph model did not directly map from the relational model, and it would not be replicable. As a result, we implemented a data migration solution rather than directly mapping the data to perform the migration on multiple occasions and promote reuse by others. We used a Jupyter Notebook project containing the relational database as CSV-files to migrate the data into the graph model format in a Neo4j database instance. Afterwards, we integrated the three external sources identified earlier with murine basal ganglia data. The source code for data integration and onboarding along with the data sets are available online[15].

From BAMS, we found brain region connectivity information possible to integrate with the murine basal ganglia data. We concluded that the retrieved BAMS connectivity information should be stored in new nodes to clarify the data's origin. We stored the BAMS brain regions in nodes with a designated node label and to presented the connectivity information through relationships between the BAMS region and basal ganglia data set's regions. To connect these regions, we performed a manual mapping between them. An expert mapped the brain regions defined in BAMS with the brain regions defined in the murine

[14] `https://neo4j.com/developer/neo4j-etl`

[15] `https://github.com/marenpg/jupyter_basal_ganglia`

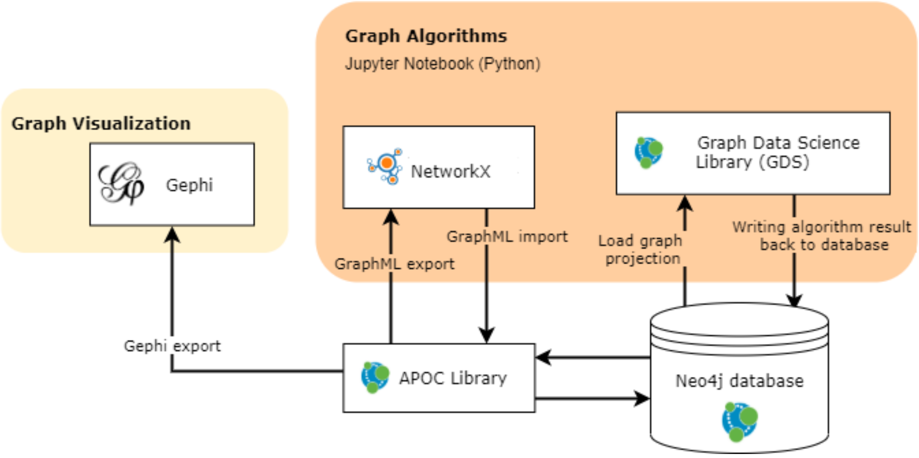


Figure 6: Solution design of the graph data analysis.

basal ganglia data set, and we stored the mapping in the data migration solution. Further, we extracted new relationships between the basal ganglia brain regions while maintaining a direct reference to the original data source. For the integration of cell descriptions from InterLex with the murine basal ganglia data set, we stored the descriptions in new nodes with a designated label. The cell types, cell groups, and cell classes in the murine basal ganglia data set contain unique identifiers from different neuroanatomical ontologies. InterLex has cell descriptions connected to multiple ontological identifiers. Based on this, we connected cell types, cell groups, and cell classes to cell descriptions based on the cell type's ontological identification attribute. The final part of the data integration extended the data set with digital cell reconstructions from NeuroMorpho.Org. NeuroMorpho.Org provides identifiers to the digital reconstructions, and some of the cell morphology nodes in the murine basal ganglia data set have an attribute for such an identifier. As with the other two initiative's data, we created a new node label to store these constructions and connect the cell morphology nodes with the digital reconstructions by matching the morphology identifiers. Figure 5 presents how data from these three sources connect to the murine basal ganglia data.

5.4 Graph analytics

We used following tools in order to realise graph analytics and visualisation and Figure 6 presents our solution for graph data analysis.

Neo4j Graph Data Science Library[16]: The Neo4j graph data science library provides a wide range of algorithms to run on projected graphs. A graph projection is a subset of the graph and can be created in Neo4j by either specifying

[16] https://neo4j.com/docs/graph-data-science/1.3

node labels and relationship labels or Cypher queries. The general process of running graph algorithms with this tool is to load the desired graph projection, run the algorithms on the projection, and finally output the result and optionally write the values back to the database. As the murine basal ganglia database is in Neo4j, Neo4j's graph data science library is a natural choice of tool to run algorithms on the data set.

NetworkX[17]: NetworkX is a Python package where one can create, manipulate, and study networks. It provides a wide range of algorithms and supports many graph file formats. Multiple Python packages provide implementations of graph algorithms. However, the Python package manager, PyPi[18], provides an API that can provide how many times Python projects have downloaded a package over the past 365 days. Searching this list for the term "network" returns the package NetworkX as the most popular. Due to its popularity and wide availability of documentation and support, it was chosen.

Gephi[19]: Gephi is an open-source software program for exploring and manipulating networks (Bastian et al. (2009)). The program provides both advanced visualisation and the possibility to manipulate the data directly in the program. In this study, the primary use of Gephi was to provide visualisations of the findings provided by the other tools. As presented above, it is possible to visualise data in both NetworkX and Neo4j. However, Gephi is very powerful in handling large amounts of data and provides multiple graph data layout algorithms. Gephi was chosen for data visualisation based on the ease of use, community support, and powerful visualisation.

5.5 Web-based access

We developed an API (available online[20]) and a Web application (application[21]and source code[22] are available online) to improve researchers' access to the data programmatically and through a user interface. The Web and API application was built based on the GRANDStack[23] architecture, which consists of GraphQL, React, Apollo, and Neo4j. This approach was chosen so that other researchers use the least possible effort to integrate the different components. GraphQL[24] is a schema-based API query language that fits well with highly interconnected data where the user of the API often needs data of multiple types simultaneously (Wieruch (2018)). As the structure allows flexible and customised queries, it is also appropriate when the use cases differ between users or are not clearly defined. React[25] is a popular JavaScript library developed by Facebook for building user interfaces (Wieruch (2018)). React is

[17] `https://networkx.org/`

[18] `https://pypi.org`

[19] `https://pypi.org/`

[20] `https://github.com/marenpg/basal_ganglia_api`

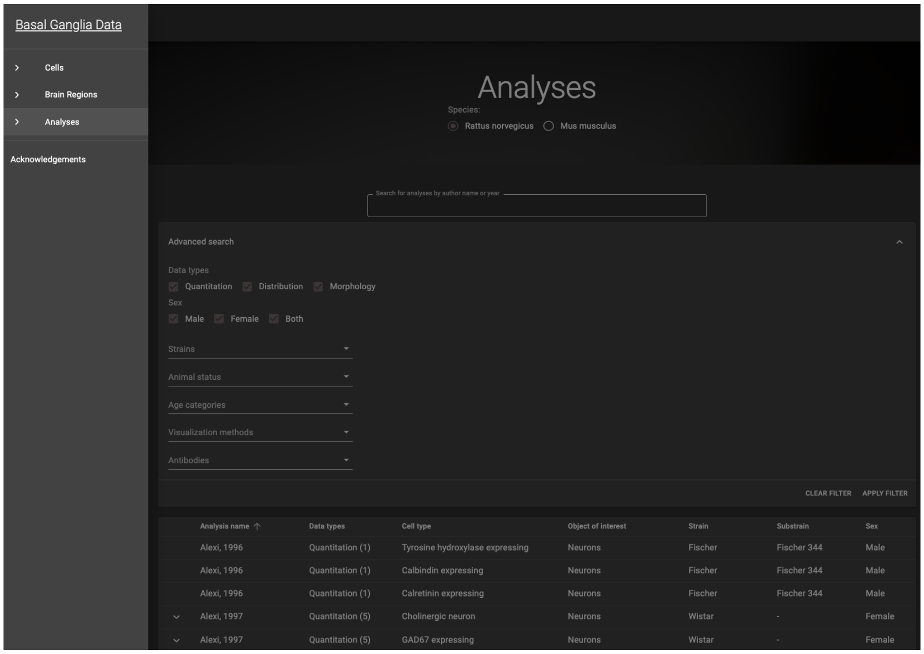
[21] `https://basal-ganglia.herokuapp.com`

[22] `https://github.com/marenpg/basal_ganglia_client`

[23] `https://grandstack.io`

[24] `https://graphql.org`

[25] `https://reactjs.org`



Species:
● Rattus norvegicus ○ Mus musculus

Data types
☑ Quantitation ☑ Distribution ☑ Morphology

Sex
☑ Male ☑ Female ☑ Both

Analysis name ↑	Data types	Cell type	Object of interest	Strain	Substrain	Sex
Alexi, 1996	Quantitation (1)	Tyrosine hydroxylase expressing	Neurons	Fischer	Fischer 344	Male
Alexi, 1996	Quantitation (1)	Calbindin expressing	Neurons	Fischer	Fischer 344	Male
Alexi, 1996	Quantitation (1)	Calretinin expressing	Neurons	Fischer	Fischer 344	Male
Alexi, 1997	Quantitation (5)	Cholinergic neuron	Neurons	Wistar	-	Female
Alexi, 1997	Quantitation (5)	GAD67 expressing	Neurons	Wistar	-	Female

Figure 7: Analyses page of the Web application.

the most popular JavaScript library per August 2020, based on downloads from the JavaScript package manager npm[26]. Apollo[27] works as the connector between the client and GraphQL API applications by offering plenty of libraries assisting effective implementation in a development stack that utilises GraphQL (Wieruch (2018)). Thus, Apollo was the natural choice for these applications.

The Web application consists of three top-level pages: one for cell types, one for brain regions, and one for analyses – see Figure 7. We also designed a page for each distinct cell type, brain region, and analysis. We chose to represent the data interconnectedness by linking the endpoints. A user can start at any entry point and find data regarding all three areas. A cell type links to brain regions and analyses, a brain region link to cell types and analyses, and a specific analysis links to at least one brain region and cell type. With these choices, the user interface followed the structure of the graph model.

The analyses page: This page displays a table of the analyses reported in the data set. On this page, the researcher can filter the results or search for analyses through a search field. The filter includes the data type, which is either a quantitation, distribution, or morphology. When a researcher selects an analysis, it opens a page with information about the selected analysis displayed in tabs. The information on the first tab differs for the three data types.

[26] `https://www.npmjs.com`

[27] `https://www.apollographql.com`

For quantitations, it presents the quantification or counting information. For example, how many investigated cell types the researcher observed in the investigated region or regions. For a distribution, the tab presents how the object of interest distributes. Finally, for morphologies, the tab presents an illustration of the cell morphology, collected from NeuroMorpho.org, together with detailed information about the investigated cell morphology. The remaining tabs on the analysis page are alike for all three data types. There is a tab with animal information that presents specimen information, including weight, age, species, and strain. The next tab displays data acquisition, presenting the research methods used to extract the analysis result, including information about the microscope used, the antibody used, and sectioning details. The anatomical metadata tab contains information about the investigated brain region and region zone and metadata about what the researchers have included in the original publication. Next, the source tab includes a reference to the original publication, including publication year and journal. Finally, there is a tab with similar analyses. The analyses presented in this tab are results from the graph data analysis.

The cell type page: On this page, the user can search for cells, or select them by their cell class or cell group, presented in a tree structure. When the researcher selects a cell type, a pane opens up and presents the cell type with a definition from InterLex, if one exists. In this pane, there are two tabs, one for brain regions and one for analyses. The brain region pane presents all the regions where experiments have recorded the cell type. Selecting a region directs the researcher to the information page of that region. In the analyses tab, the researcher can see all analyses that investigate the selected cell type. Selecting an analysis from the list navigates the user to the information page of that analysis.

The brain regions page: This page presents all regions related to the basal ganglia for mice and rats in two tree structures. The researcher can search for a region and filter on species. Selecting a region opens up a side pane like with the cell types. There are two tabs for the mouse brain regions and three for the rat brain regions in this pane. The third tab presents connectivity information derived from BAMS. In this tab, the user can see the regions connected to the selected region and filter on direction and strength. Selecting a connection displays the original connectivity information from BAMS with citations and links to where we have collected the connection. The two other tabs are cell types and analyses. The analyses tab displays a list of the analyses that have investigated the selected brain region. The cell type tab presents all the cell types that experiments have recorded in the selected region.

6 Evaluation

We successfully integrated the relevant data sets increased the connections to the three primary access nodes: cell types, brain regions, and analyses. Compared to the original database structure, the graph model presents higher connectivity for these nodes. The database generated in this study consists of 9539 distinct nodes with 46 distinct node labels, 29807 distinct relationships, and 66 distinct relationship types. Further, we extended the data set with 142 nodes with three labels from integration with external sources. Eighty of these nodes were brain regions from BAMS. The integration added 351 new distinct relationships, where 335 of these represent brain region connectivity. This shows that a graph-based representation of brain-related data facilitate the integration of data from a variety of neuroscience data sets. In what follows, we present:

1. The results of a set of data analyses we conducted on the resulting integrated graph in order to demonstrate that a graph-based approach could help gaining a better understanding of the data for brain-related data sets; and

2. The results of the user study for the Web access interface built on top of the graph model in order to show that a graph-based data organisation could improve accessibility of the brain-related data.

6.1 Data analysis

We considered data analysis from exploratory data analysis and confirmatory data analysis perspectives (Hartwig (1979)). The former aims at obtaining general information about the data and the latter to answer specific questions, while the latter is concerned with proving or answering a specific hypothesis or question. We developed a Jupyter Notebook project containing the complete algorithm set-up for the graph data analysis and made it available online[28].

We first visualised the entire graph to see if we could observe any clustering. To visualise the entire graph, we loaded all the nodes and relationships into the visualisation tool Gephi. Figure 8 presents this visualisation which uses the ForcedAtlas2 graph layout algorithm (Jacomy et al. (2014)). From the entire graph visualisation, we observed that the data naturally groups into two almost separate clusters. There is a large cluster on the right side and a smaller cluster on the left side, with some nodes combining them. Investigating the smaller cluster, we identify that it solely consists of data concerning considered papers and their exclusion reason. It is only source nodes that combine the two clusters. We utilised community detection algorithms to investigate the graph data structure, specifically the Label propagation algorithm (LPA) and Louvain algorithm. We chose these algorithms as the LPA performs community detection based on the structure, while the Louvain algorithm applies heuristics based on the nodes' modularity. A suitable method for finding influential nodes in a graph is to run centrality algorithms. There are multiple centrality algorithms. We applied the PageRank algorithm (e.g., Figure 9) and betweenness centrality algorithm in Neo4j, and the closeness centrality, the betweenness centrality, and the HITS algorithm from NetworkX. We chose these algorithms as they implement differing measures for centrality, including direct and indirect influence. We used a node Similarity algorithm through the Neo4j GDS library to analyse

[28] `https://github.com/marenpg/jupyter_basal_ganglia`

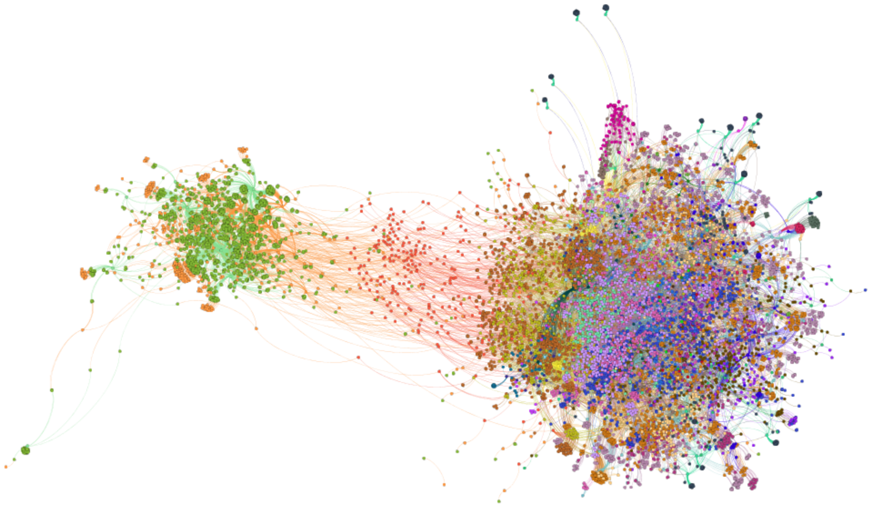


Figure 8: The murine basal ganglia data set visualised using the ForcedAtlas2 layout algorithm in Gephi.

similarities. We chose to use the Neo4j Node Similarity algorithm mainly due to its efficiency for comparing all of the graph's nodes. Figure 9 presents the nodes representing sexes and analyses and their relationships. The middle large cluster represents males, the bottom cluster represents females, and the top cluster represents both sexes. The neuroscience expert stated that it is common knowledge that most research uses male specimens.

We evaluated the results with an expert and Table 1 summarises the findings and evaluations from the exploratory graph data analysis. In summary, the information extracted with graph data analysis provided a good understanding of the structure and content of the data set.

Result - information	**Algorithms**	**Evaluation**
Cell type information: The most investigated cell type category is "Expressing"	Louvain, LPA, PageRank	This is already known.
Data structure: Community around the data set species.	Louvain, LPA	This is already known.
Data quality: Chemical solution *Unspecified* is the second most used anesthetic and fourth most used perfusion fix medium.	PageRank, Louvain	Expected as it is known that solution information is poorly reported but fascinating to observe for this data.

Data quality: The second most used software is *Unspecified*.	PageRank	Unexpected and interesting as it will make the research results challenging to reuse.
Source information: The most influential publications are *Neuroscience*, *Brain Research*, and *Journal of Comparative Neurology*.	PageRank	Already known as the original database paper by Bjerke et al. also stated this.
Cell type information: *Neuron* is the most investigated cell type.	PageRank, Closeness centrality	Expected as it is the easiest for scientists to observe.
Cell type information: *Tyrosine hydroxylase expressing* (TH) cells are the second most investigated cell type.	PageRank, Closeness centrality	Already known due to TH-cells relevance in Parkinson's disease.
Cell type information: Most of the experiment investigate entire neurons.	PageRank, HITS	Interesting as it implies that neuroscience knows much more about the whole cells than the sub-cellular entities.
Method information: "Bright-field microscope" is the most used microscope type.	PageRank	Expected as it is a very common microscope.
Method information: Immunohistochemistry is the most used visualisation method, histochemistry the second most. No difference between species.	PageRank	Expected based on the data Bjerke at al. collected.
Method information: Tyrosine hydroxylase and Rabbit antibody are the most used reporter targets.	PageRank	Expected as the data contains many TH studies.
Method information: "Goat anti rabbit_biotin" is the most used Reporter.	PageRank, HITS	Unexpected and interesting as it can tell us something about the data quality.
Method information: The most used sectioning instrument is "Cryostat", followed closely by "Freezing microtome".	PageRank	Not evident but of little interest as cutting instruments are mundane.
Brain region information: The data set contains the most information about the brain region caudoputamen for both species.	PageRank, Betweenness centrality, Node similarity	Already known from the data set paper.

Brain region information: Most of the data investigates the rostral region zone.	Betweenness centrality	Expected, but interesting as it displays a bias in the data.
Brain region information: When a study investigates the internal segment region, they also investigate the external segment region.	Node similarity	Already known as they are sub-regions of the same region.
Specimen information: “Adult” is the age category, most often used (big difference).	PageRank	Expected as researchers use other age categories mostly for research specific to the age category.
Specimen information: Most influential strain is Wistar for rats, $C57BL6$ for mice.	PageRank	Expected as these are common strains.
Specimen information: “Male” is the most influential sex.	PageRank	Expected, but interesting as it displays a bias in the research.

Table 1: Graph data analysis results and evaluation.

In the confirmatory data analysis part, we aimed to find similar analyses based on a specific criteria. The analysis nodes represent one of the data sets’ three primary entry points and are what researchers often use when comparing results. We were interested in finding analyses investigating the same cell type in the same brain region and having the same object of interest. As with the exploratory data analysis, we utilised the Neo4j implementation of the Node Similarity algorithm. Compared with the exploratory data analysis, we used a slightly adjusted graph projection because we only want the analyses that were entirely similar with respect to cell type, brain region, and object of interest. We created a graph projection containing only the four relevant labels with a direct relationship between them. The node similarity algorithm ran on this projection with degree-cutoff set to 3 and similarity-cutoff set to 1 and configured to write the relationship back to the graph for the nodes that matched the criteria. These efforts created a relationship between the analyses with the same cell type, brain region, and object of interest. Figure 10 presents the analyses (in orange) in the data set connected to the specified nodes and species. The yellow nodes represent the two species in the data set, and the central node in the middle is the cell type “neurons”.

We evaluated the result by querying nodes and nodes found similar and manually verified that they match the requirements presented and the results are correct. We managed to find similar analyses successfully with a few lines of code.

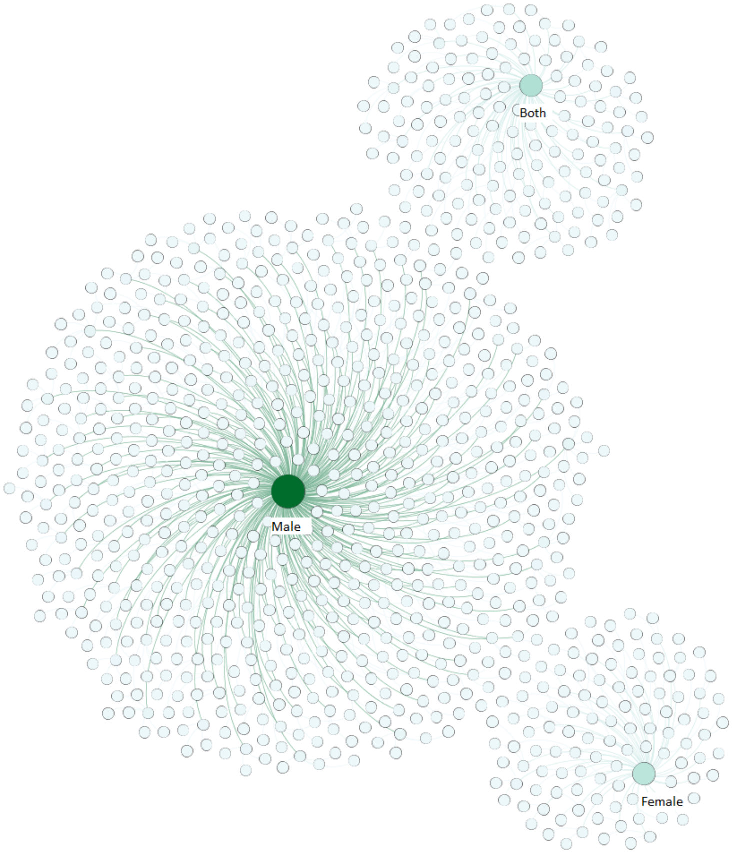


Figure 9: Graph-visualisation of analyses and the sex they study from the murine basal ganglia data set.

6.2 User study

We applied formative testing for testing the usability of the Web interface. The following list presents the steps of the usability testing performed:

1. The observer introduces the study and the Web application.
2. The observer describes how to think-aloud and encourages the user to apply the technique.
3. The observer presents the tasks to the user and explains that there will be no communication during the tasks, and if the user can not complete a task, they should continue with the next.
4. The user performs the tasks while thinking out loud.

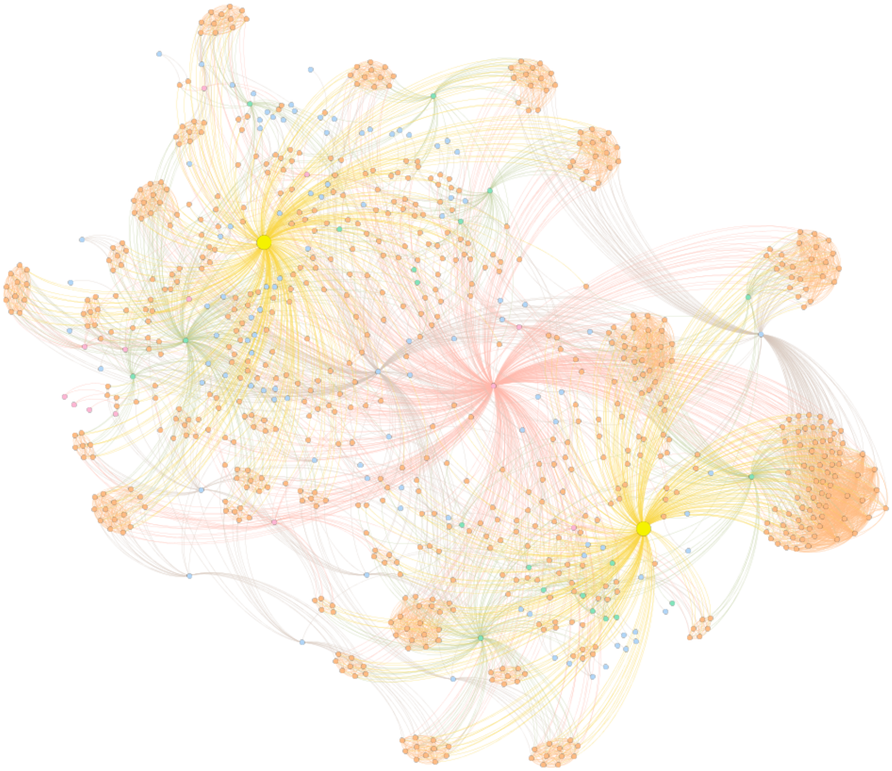


Figure 10: The data set analyses with related nodes.

5. The observer interviews the user to evaluate the user's overall experience.

We contacted neuroscience and medicine faculties from multiple universities. However, all the researchers who expressed interest in participating were from the University of Oslo. We verified that the participants were researchers, often working with publicly available data. The test was executed remotely using an online conferencing solution. Between each participant, we performed small adjustments to the applications according to feedback. The users completed the tasks sequentially. There were total of 4 participants and 28 tasks (leading to 112 observations) performed by the participants. The tasks are listed in Table 2.

#	**Task**	**Type**
1a	Can you find how many analyses have been performed on Rattus norvegicus?	Task
1b	Can you find how many analyses have been performed on Mus musculus?	Task

2	How many morphology analyses have been performed on the species Mus musculus?	Task
3	How many analyses, performed on rat, have used the antibody with unique id RRID:AB_476894	Task
4	How many analyses are performed on a juvenile rat (19-28 days)?	Task
5	Can you find the study by Fujiyama (2016) among the analyses?	Task
5a	In the study by Fujiyama (2016), how many axonal varicosities in total were observed in the substantia nigra?	Sub-task
5b	In this study, what was the weight range of the specimens used?	Sub-task
5c	In which journal was this study published?	Sub-task
6	See if you can find the study by Echeverry (2004) on NADPHD expressing neurons?.	Task
6a	In this study, what part of the Caudoputamen was covered?	Sub-task
7	How many analyses that have been performed on the rat substantia nigra?	Task
8	Can you find the number of regions that are connected to the rat Caudoputamen?	Task
8a	Which of the connected regions does the Caudoputamen have a very strong, afferent relationship to?	Task
8b	Can you find how these relationships were derived?	Sub-task
9	In the mouse Caudoputamen, how many mixed class neuron cell types are observed?	Task
10	For this region, can you find how many analyses are performed on dopamine 1 receptor expressing cells?	Task
11	Staying on this page, can you get back all the analyses performed on mus musculus?	Task
12	See if you can find a morphology analysis of medium spiny neuron cells.	Task
12a	Select one of these morphologies.	Sub-task
12b	From which repository was the morphology illustration collected?	Sub-task
13	How many cells are returned when searching for “dopamine receptor”?	Task
14	Can you find a description of the cell type “Glia”?	Task

14a	Where was this description collected from?	Sub-task
15	Calretinin expressing interneuron is the cell type investigated in a number of analyses, can you find how many?	Task
15a	Can you find how many brain regions Calretinin expressing interneuron are observed in?	Sub-task
15b	Can you find the number of analyses concerning Calretinin expressing interneuron in the substantia nigra?	Sub-task
16	Can you find the sources and repositories that have contributed to the website data	Task

Table 2: The tasks used in the usability study.

The first participant failed in one task (Task 9), partially completed two (Tasks 5a and 12b), and completed the rest successfully. The second participant completed two tasks partially (Tasks 5a and 9), and the rest fully, while the third participant completed two tasks partially (Tasks 2 and 6) and rest successfully. Finally, the last participant completed all the tasks successfully. The first two participants struggled with task 5a and task 9, while the third participant experienced some struggles with the filter function. Task 5.1 regards finding the total number of axonal varicosities observed by Fujiyama in 2016 in the substantia nigra. After observing the two first participants struggling with finding this number, we adjusted the interface to present this number more clearly, and according to the final usability tests, this was successful. Task 9 asks the user to find the number of mixed-class neurons observed in the mouse caudoputamen. When the user selects a cell type, the resulting page presents all the cell types observed in that region. The goal was that the user should count the number of mixed-class neuron cell types on this page. For the first two participants, this was not clear. The first participant also struggled with Task 12, where the user was to find a morphology illustration's source repository. We updated the page to reference the morphology repository more clearly, and the next participants found it with ease.

The third participant struggled with the filter-function at the beginning of the test, which caused the unsatisfactory completion of Task 2 and Task 6. However, the participant learned how it worked and managed all the subsequent tasks. In summary, the user feedback improved the applications to a point where the users managed to complete almost all tasks confidently. Further, the participants grew more confident throughout the usability test. In the user interviews, performed right after the tasks, we asked the users about their overall user experience. All the participants had an overall good impression. They felt they understood the application and that the interface provided the necessary entry points for finding data relevant to them. One participant suggested the possibility for community building, such as having a contact page with more information and sharing data.

7 Conclusions

In this study, we presented a graph-based approach for representing neuroscience data, exemplified with the murine basal ganglia data set. We addressed multiple ways of working with a graph model in the neuroscience domain from a data management perspective based on the proposed data set graph model. The study described how data from external sources can integrate with neuroscience data in a graph model, applications for Web-based access to improve the usability of the data, and the use of graph analytics to extract new information and improve the understanding of the data. Further, the study presented evaluations of the developed software, the usability of the data, and the results obtained by applying graph algorithms. We presented definite advantages of graph-based data representation through our work, including ease of data analysis, support for data integration, and availability through Web-based data access. Many of the presented areas of graph-based data representation in the neuroscience domain are still uncharted terrain. It is relevant to continue evaluating the implications of graph-based data representation and work to solve the challenges with data management in the field of neuroscience.

Regarding the future work, firstly, while extending the murine basal ganglia graph model with data from other neuroscience data initiatives, it was challenging to obtain information about the content the initiatives provided, programmatic data access, and, occasionally, the data format. We suggest that further research performs a thorough review of neuroscience data initiatives and present what data are available from where and how the researchers can access the data, preferably including the data formats. Another approach could be to look at standardisation for programmatic access to neuroscience data. Secondly, further research can investigate a graph-based approach for representing other data types (Reutter (2020)) to observe if the benefits and challenges are different for these and on a larger data set in combination with other graph analytics techniques to evaluate the scalability related to querying performance (da Silva & Maia (2019), Tjendry & Istiono (2020)) and usability (Soylu et al. (2018)) related to data integration, access, and analytic processes.

Acknowledgements

This work was partly funded by the EC H2020 DataCloud project (Grant number 101016835).

References

Akil, H., Martone, M. E. & Van Essen, D. C. (2011), 'Challenges and Opportunities in Mining Neuroscience Data', *Science* **331**(6018), 708–712.

Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A. & Lippert, T. (2016), 'The human brain project: creating a European research infrastructure to decode the human brain', *Neuron* **92**(3), 574–581.

Angles, R. & Gutierrez, C. (2008), 'Survey of graph database models', *ACM Computing Surveys* **40**(1), 1–39.

Ascoli, G. A. (2012), 'Mobilizing the base of neuroscience data: the case of neuronal morphologies', *Nature Reviews Neuroscience* **4**, 318–324.

Bassett, D. S., Zurn, P. & Gold, J. I. (2018), 'On the nature and use of models in network neuroscience', *Nature Reviews Neuroscience* **19**, 566–578.

Bastian, M., Heymann, S. & Jacomy, M. (2009), Gephi: an open source software for exploring and manipulating networks, *in* 'Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM 2009)', pp. 361–362.

Bjaalie, J. G. (2002), 'Localization in the brain: new solutions emerging', *Nature Reviews Neuroscience* **3**(4), 322–325.

Bjerke, I. E., Øvsthus, M., Andersson, K. A., Blixhavn, C. H., Kleven, H., Yates, S. C., Puchades, M. A., Bjaalie, J. G. & Leergaard, T. B. (2018), 'Navigating the murine brain: toward best practices for determining and documenting neuroanatomical locations in experimental studies', *Frontiers in Neuroanatomy* **12**, 82.

Bjerke, I. E., Puchades, M. A., Bjaalie, J. G. & Leergaard, T. (2020), 'Database of literature derived cellular measurements from the murine basal ganglia', *Scientific data* **7**(1), 1–14.

Bjerke, I. E., Puchades, M. A., Bjaalie, J. G. & Leergaard, T. B. (2019), 'Database of quantitative cellular and subcellular morphological properties from rat and mouse basal ganglia [Data set]', *Human Brain Project Neuroinformatics Platform* .

Bota, M. & Swanson, L. W. (2010), 'Collating and curating neuroanatomical nomenclatures: principles and use of the Brain Architecture Knowledge Management System (BAMS)', *Frontiers in Neuroinformatics* **4**, 3.

Bowden, D. M., Song, E., Kosheleva, J. & Dubach, M. F. (2012), 'NeuroNames: an ontology for the BrainInfo portal to neuroscience on the web', *Neuroinformatics* **10**(1), 97–114.

Bunner, K. D. & Rebec, G. V. (2016), 'Corticostriatal dysfunction in Huntington's disease: the basics', *Frontiers in Human Neuroscience* **10**, 317.

Campbell, N. A., Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V. & Jackson, R. B. (2011), *Biology*, San Francisco, CA: Pearson Benjamin Cummings.

Cattell, R. (2011), 'Scalable SQL and NoSQL data stores', *Sigmod Record* **39**(4), 12–27.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2009), *Introduction to algorithms*, MIT press.

da Silva, F. T. & Maia, J. E. B. (2019), 'Query Expansion in Text Information Retrieval with Local Context and Distributional Model', *Journal of Digital Information Management* **17**(6), 313–320.

Diestel, R. (2017), *Graph Theory*, Vol. 173 of *Graduate Texts in Mathematics*, 5 edn, Springer Berlin Heidelberg.

Drachman, D. A. (2005), 'Do we have brain to spare?', *Neurology* **64**(12), 2004–2005.

EBRAINS (n.d.), `https://ebrains.eu/`. Last accessed: 2020-11-04.

Fan, X. & Markram, H. (2019), 'A Brief History of Simulation Neuroscience', *Frontiers in Neuroinformatics* **13**, 32.

Fernandes, D. & Bernardino, J. (2018), Graph databases comparison: Allegrograph, arangoDB, infinitegraph, Neo4J, and orientDB, *in* ‘Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)’, SciTePress, pp. 373–380.

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., Halavi, M., Kennedy, D. N., Marenco, L., Martone, M. E., Miller, P. L., Müller, H.-M., Robert, A., Shepherd, G. M., Sternberg, P. W., Van Essen, D. C. & Williams, R. W. (2012), ‘The Neuroscience Information Framework: A Data and Knowledge Environment for Neuroscience’, *Neuroinformatics* **6**, 149–160.

Gerfen, C. R. & Bolam, J. P. (2016), The neuroanatomical organization of the basal ganglia, *in* ‘Handbook of Behavioral Neuroscience’, Vol. 24, Elsevier, pp. 3–32.

Hamilton, D., Shepherd, G., Martone, M. & Ascoli, G. (2012), ‘An ontological approach to describing neurons and their relationships’, *Frontiers in Neuroinformatics* **6**, 15.

Han, J., Haihong, E., Le, G. & Du, J. (2011), Survey on NoSQL database, *in* ‘Proceedings of the 6th International Conference on Pervasive Computing and Applications (ICPCA 2011)’, IEEE, pp. 363–366.

Hartwig, F. (1979), *Exploratory data analysis*, Quantitative applications in the social sciences, SAGE.

Hecht, R. & Jablonski, S. (2011), NoSQL evaluation: A use case oriented survey, *in* ‘Proceedings of the 2011 International Conference on Cloud and Service Computing (CSC 2011)’, IEEE, pp. 336–341.

Herculano-Houzel, S. (2009), ‘The human brain in numbers: a linearly scaled-up primate brain’, *Frontiers in Human Neuroscience* **3**, 31.

Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. (2014), ‘ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software’, *PloS one* **9**(6), 1–12.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., of Biochemistry, D., Jessell, M. B. T., Siegelbaum, S. & Hudspeth, A. (2000), *Principles of neural science*, 5 edn, McGraw-hill New York.

KnowledgeSpace (n.d.), `https://knowledge-space.org/about`. Last accessed: 2020-10-30.

Merriam-Webster.com dictionary (n.d.), ‘Nomenclature’, https://www.merriam-webster.com/dictionary/nomenclature. Online; accessed 7 December 2020.

Middleton, F. A. & Strick, P. L. (2000), ‘Basal ganglia and cerebellar loops: motor and cognitive circuits’, *Brain Research Reviews* **31**(2-3), 236–250.

Miller, E. (1998), ‘An introduction to the resource description framework’, *Bulletin of the American Society for Information Science and Technology* **25**(1), 15–19.

National Academies of Sciences, Engineering, and Medicine (2020), *Neuroscience Data in the Cloud: Opportunities and Challenges: Proceedings of a Workshop*, The National Academies Press, Washington, DC.

Needham, M. & Hodler, A. E. (2019), *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*, O'Reilly Media.

Newman, M. (2018), *Networks*, 2 edn, Oxford university press.

Obeso, J. A., Marin, C., Rodriguez-Oroz, C., Blesa, J., Benitez-Temiño, B., Mena-Segovia, J., Rodríguez, M. & Olanow, C. W. (2008), 'The basal ganglia in Parkinson's disease: current concepts and unexplained observations', *Annals of Neurology* **64**(S2), S30–S46.

Özsu, M. T. & Valduriez, P. (2019), *Principles of distributed database systems*, 4 edn, Springer.

Petilla Interneuron Nomenclature Group (PING) (2008), 'Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex', *Nature Reviews Neuroscience* **9**(7), 557–568.

Pérez, J., Arenas, M. & Gutierrez, C. (2009), 'Semantics and complexity of SPARQL', *ACM Transactions on Database Systems* **34**(3), 30–43.

Reutter, J. L. (2020), Current Challenges in Graph Databases (Invited Talk), *in* 'Proceedings of the 23rd International Conference on Database Theory (ICDT 2020)', pp. 3:1–3:1.

Robinson, I., Webber, J. & Eifrem, E. (2013), *Graph databases*, O'Reilly Media, Inc.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. (2008), 'The graph neural network model', *IEEE Transactions on Neural Networks* **20**(1), 61–80.

Shepherd, G. M., Marenco, L., Hines, M. L., Migliore, M., McDougal, R. A., Carnevale, N. T., Newton, A. J., Surles-Zeigler, M. & Ascoli, G. A. (2019), 'Neuron Names: A Gene- and Property-Based Name Format, With Special Reference to Cortical Neurons', *Frontiers in Neuroanatomy* **13**, 25.

Sicilia, M.-A., García-Barriocanal, E. & Sánchez-Alonso, S. (2017), 'Community curation in open dataset repositories: Insights from Zenodo', *Procedia Computer Science* **106**, 54–60.

Sivagnanam, S., Majumdar, A., Yoshimoto, K., Astakhov, V., Bandrowski, A., Martone, M. E. & Carnevale, N. T. (2013), Introducing the Neuroscience Gateway, *in* 'Proceedings of the 5th International Workshop on Science Gateways (IWSG 2013)', Vol. 993 of *CEUR Workshop Proceedings*, CEUR-WS.org.
URL: *http://ceur-ws.org/Vol-993/paper10.pdf*

Soylu, A., Kharlamov, E., Zheleznyakov, D., Jiménez-Ruiz, E., Giese, M., Skjæveland, M. G., Hovland, D., Schlatte, R., Brandt, S., Lie, H. & Horrocks, I. (2018), 'OptiqueVQS: A visual query system over ontologies for industry', *Semantic Web* **9**(5), 627–660.

Swanson, L. W. (2000), 'What is the brain?', *Trends in Neurosciences* **23**(11), 519–527.

Teeters, J. L., Harris, K. D., Millman, K. J., Olshausen, B. A. & Sommer, F. T. (2008), 'Data Sharing for Computational Neuroscience', *Neuroinformatics* **6**(1), 47–55.

Tjendry, D. & Istiono, W. (2020), 'Is the Binary Search Faster when Two Variables are Added in the Middle of the Data?', *Journal of Digital Information Management* **18**(2), 57–64.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H. et al. (2013), 'The WU-Minn human connectome project: an overview', *Neuroimage* **80**, 62–79.

Wieruch, R. (2018), *The Road to GraphQL: Your journey to master pragmatic GraphQL in JavaScript with React. js and Node. js*, Robin Wieruch.

Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. & Sun, M. (2018), 'Graph Neural Networks: A Review of Methods and Applications', *arXiv* .
URL: *https://arxiv.org/abs/1812.08434*