Plan Overview

A Data Management Plan created using DMPonline

Title: Enhancing Software for the SpiNNaker Community

Creator: Andrew Rowley

Principal Investigator: Andrew Rowley

Affiliation: University of Manchester

Funder: Engineering and Physical Sciences Research Council (EPSRC)

Template: EPSRC Data Management Plan Customised By: University of Manchester

Project abstract:

SpiNNaker-2, a new version of the SpiNNaker million-core neuromorphic computing platform, has been created. The software written for the existing platform allows users to describe, execute, and obtain results from the simulation of biological-scale neural networks in real-time and connect these to external robotic devices. This software will be updated to allow the execution of these networks on and connection of these devices to the SpiNNaker-2 platform, enabling research into robotics and artificial intelligence to continue into the future.

ID: 84403

Start date: 01-10-2022

End date: 30-09-2024

Last modified: 29-09-2021

Enhancing Software for the SpiNNaker Community

Manchester Data Management Outline

1. Will this project be reviewed by any of the following bodies (please select all that apply)?

- Funder

2. Is The University of Manchester collaborating with other institutions on this project?

- No – only institution involved

3. What data will you use in this project (please select all that apply)?

- Acquire new data
- Re-use existing data (please list below)

Reports, Software and Documentation will be created. The starting point will be existing Software and Documentation. Users using the software might end up generating their own data which will be stored on services managed by the project.

4. Where will the data be stored and backed-up during the project lifetime?

- Other storage system (please list below)

Software will be stored on GitHub; as git is a distributed software system this will also be mirrored by anyone who clones the repository. It will also be stored in publishing services such as Zenodo.
Documentation will be created from the source code, so will be stored in the same place.
Any user-generated data will only be stored temporarily until users extract the data for their own use.

5. If you will be using Research Data Storage, how much storage will you require?

- Not applicable

6. Are you going to be receiving data from, or sharing data with an external third party?

- Yes

Users may submit data to be processed by the services provided by the project. Users may then download their results data after processing.

7. How long do you intend to keep your data for after the end of your project (in years)?

- 21+ years

Data will be stored only in external services; it is unknown how long GitHub will archive projects for. Similarly, Zenodo will hold copies of the software, which are assumed to exist in perpituity.

Guidance for questions 8 to 13

Highly restricted information defined in the Information security classification, ownership and secure information handling SOP is information that requires enhanced security as unauthorised disclosure could cause significant harm to individuals or to the University and its ambitions in respect of its purpose, vision and values. This could be: information that is subject to export controls; valuable intellectual property; security sensitive material or research in key industrial fields at particular risk of being targeted by foreign states. See more examples of highly restricted information .

Personal information, also known as personal data, relates to identifiable living individuals. Personal data is classed as special category personal data if it includes any of the following types of information about an identifiable living individual: racial or ethnic origin; political opinions; religious or similar philosophical beliefs; trade union membership; genetic data; biometric data; health data; sexual life; sexual orientation.

Please note that in line with data protection law (the UK General Data Protection Regulation and Data Protection Act 2018), personal information should only be stored in an identifiable form for as long as is necessary for the project; it should be pseudonymised (partially de-identified) and/or anonymised (completely de—identified) as soon as practically possible. You must obtain the appropriate ethical approval in order to use identifiable personal data.

8. What type of information will you be processing (please select all that apply)?

- No confidential or personal data

9. How do you plan to store, protect and ensure confidentiality of any highly restricted data or personal data (please select all that apply)?

- Not applicable

10. If you are storing personal information (including contact details) will you need to keep

it beyond the end of the project?

- Not applicable

11. Will the participants' information (personal and/or sensitive) be shared with or accessed by anyone outside of the University of Manchester?

- Not applicable

12. If you will be sharing personal information outside of the University of Manchester will the individual or organisation you are sharing with be outside the EEA?

- Not applicable

13. Are you planning to use the personal information for future purposes such as research?

- No

14. Will this project use innovative technologies to collect or process data?

- Yes, and innovative technologies will not collect or process personal data (please list the innovative technologies below)

The SpiNNaker 1 and SpiNNaker 2 Neuromorphic platforms will be used to process user data, but this won't include personal data.

15. Who will act as the data custodian for this study, and so be responsible for the information involved?

Andrew Rowley

16. Please provide the date on which this plan was last reviewed (dd/mm/yyyy).

2021-09-17

Data Collection

What data will you collect or create?

- Outputs of the execution of neural networks. These are binary-formatted datasets that users can choose to save during the execution of their network. The data is generated using the Neo library (https://neo.readthedocs.io/en/stable/), which supports a wide range of data formats that are commonly used in the research communities supported, as well as other non-proprietary formats. The size of the data varies depending on the neural networks being used, but these can grow to sizes in GB or TB for large networks.
- Software source code. This is the code that is used to set up and run the neural networks on SpiNNaker. The source code is written in Python, Java, C and ARM assembly code. The size of this data is generally of the order of at most a few GB, though it is generally less than this.
- Software releases. The software is release is made up of many parts. The C and ARM assembly code is compiled into SpiNNaker APLX files (https://spinnakermanchester.github.io/docs/spinn-app-3.pdf). These are then embedded in the Python source tree. The Python code is packaged as-is. The Java code is compiled into Jar files. Software releases are at most several MBs in size.
- Software documentation. This is the user and system documentation of the software. Some documentation is stored in the code itself (inline documentation), but there are also a number of documents stored in GitHub Markdown files and in Google Docs format. These are sized in the MBs.

How will the data be collected or created?

Data will be generated when a user runs a neural network and chooses to record aspects of that execution. Data will be generated using the Neo library (https://neo.readthedocs.io/en/stable/), which can generate data in a number of formats depending on the needs of the users. This library already includes suitable documentation for users to install, run and develop the software as needed, which will then ensure that the data generated is consistent over multiple runs. When the user uses either the SpiNNaker Jupyter service, or their own machines, they are responsible for their own data storage, organisation and versioning. When using our batch execution service, the data will be stored in local folders on the service machine and held temporaily for up to 3 months or until deleted by the user. This data is organised into folders which are named according to the identifier of the job they are executing, and the folder timestamps are used to identify the data creation date. It is not possible to overwrite this data using the service, so versioning is not implemented here.

Software source code will be written on local developer machines, and make use of git hosted on an external GitHub service. Developers are encouraged to commit and push changes often for purposes of collaboration and to ensure changes to the software are saved in more than one place. This service is well documented, and files are organised into various modules and packages. Consistency of the data is ensured over time through automatic style checking, which ensures that code remains readable and maintainable over time. Versioning is provided by git.

Software releases are generated at some time during development, when the tests of the software are passing, and the software is stable. Software releases are stored in external software repositories; Python releases are stored in PyPI and Java releases are stored in Maven. These modules already contain the appropriate metadata, which is updated during generation of the release. This is enforced by the external services which require metadata to be present. The release process is also documented to ensure this are consistent across releases.

Software documentation is generated when code is merged into the master / main branch. The software documentation is tested to ensure consistency over different versions. The documentation is automatically pushed to the GitHub pages sites using GitHub actions.

Documentation and Metadata

What documentation and metadata will accompany the data?

The data format contains metadata regarding the contents of the data. The data itself is generated for the user, so it will be up to the user to control any additional metadata in relation to the data. The neo library is a candidate standard in itself (https://www.incf.org/sbp/neo).

The software code itself will be accompanied by metadata in the form of tags and README documents on the GitHub service, which will then allow the searching of the software. The software documentation will also be linked to the repository to allow others to make use of it.

The software releases are accompanied by metadata that is enforced by the repositories on which they are hosted. Python software follows the Python packaging metadata standards (https://packaging.python.org/specifications/core-metadata/) and Java follows the Maven POM metadata standard (https://maven.apache.org/ref/3.8.2/maven-model/maven.html).

Ethics and Legal Compliance

How will you manage any ethical issues?

Data is fully in control of the user who generated it, and that user can therefore control the data in a way that makes sense for their own project.

We are aware of potential dual-use-of-concern issues surrounding artificial intelligence applications, and the PI has been involved in ethics discussions in the Human Brain Project (https://www.humanbrainproject.eu/en/social-ethical-reflective/about/ethics-rapporteur-programme/). We will continue to monitor the potential uses of the software and ensure that users of the services are asked that they will not make use of the services for such uses, and no development will be done or support will be given where such use is apparent.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Users are in full control of their own data, and other users of the service cannot access each others data. This is protected by standard Linux access mechanisms.

Software stored on GitHub is licensed appropriately and under the copyright of the developers. The University of Manchester is expected to retain the copyright and IPR associated with any developments done as part of this project.

Storage and Backup

How will the data be stored and backed up during the research?

Any user-generated data will be stored on local server hard disks, which are designed for high-availability through the use of RAID arrays, which are monitored for errors. The local servers in question are hosted in the SpiNNaker machine room to ensure they are as close to the hardware as possible. As such they do not make use of institutional storage services. This is also not required in any case; as user data is only expected to exist for relatively short periods of time on the local service, no backup of this data is provided.

Software source code is stored externally on GitHub. As git is a distributed system, any user that clones any of the software will also create a backup of the software at that point. They are additionally put on Zenodo to allow referencing of the code. Additional local backups of storage in Manchester is also done.

Software releases are stored on external repository services. These are not backed up explicitly, though these services do have some backups themselves. Releases can always be generated from the source code in any case, so such backups are not required.

How will you manage access and security?

Access to the services is currently through externally managed OIDC services. No local management of access is therefore required.

The software is stored externally on GitHub, where access can be managed at the organisational or user level. The updating of the access rules is a task that will continue to be undertaken by the team responsible for the software.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

All data on the service is expected to be held only for short periods of time. The users are responsible for retaining their own data and sharing it outside of any service run within this project.

The software is expected to be stored on GitHub and/or Zenodo for as long as such services continue to exist.

What is the long-term preservation plan for the dataset?

Users are responsible for their own data preservation.

The software is expected to be stored on GitHub and/or Zenodo for as long as such services exist. Manchester institutional data repositories can also be used should these external services cease to exist.

Data Sharing

How will you share the data?

The data generated by users in this project is not part of the project.
Software written in this project will extend existing open source software and will remain open source beyond the end of the project. The software is always openly available via GitHub, with releases being available on Zenodo, PyPI and Maven as appropriate.

Are any restrictions on data sharing required?

Software is currently under a GPLv3 license, which only restricts certain commercial activities. As this is entirely owned by the University of Manchester, it could be relicensed as appropriate in the future should this be required.

Responsibilities and Resources

Who will be responsible for data management?

Principle: Andrew Rowley, Senior Research Software Engineer.
Other members of the SpiNNaker team are also responsible for source code and releases. These members are not explictly named as the makeup of the team may change over the course of the project.

What resources will you require to deliver your plan?

Standard existing computers and external services will be used.