Grant Number:	720270	Grant Title:	Human Brain Project SGA1

Deliverable Title:	D1.5.3 – Detailed plan of data usage and the impact of generated data on models
Contractual Number and type:	SGA1 D1.5.3 - Report (includes SP1 Data Management Plan HBP-SGA1-SP1DMP-M23-v1.3)
Dissemination Level:	PU (= Public)
Version / Date:	V4.5 - 20 July 2018 Resubmitted: 30 July 2018; Accepted: 7 Dec 2018

Abstract:	This Deliverable is SP1's Detailed plan of data usage and the impact of the generated data on models. During SGA1, SP1 has been structured in order to provide the critical data required for modelling (reconstruction and validation) of the 4 main brain subcircuits (cerebral cortex, cerebellum basal ganglia and hippocampus) at the cellular, subcellular, and network level. Conversely, SP1 was not intended to provide full coverage of data for whole-brain reconstruction. From M1 to M24 of SGA1, a detailed plan for data usage via cross-SP working meetings was drawn up as stated in the DoA. This plan includes strategic structural and functional data to model and simulate the four major brain circuits and to set up their use in simulation/modelling at sub-cellular, cellular and circuit levels. Due to time and resource pressures, the data generated was sent to the Platforms as it became available, i.e., not only the initial and intermediate datasets, but also incomplete ones, to test all the pipelines. Identification of the gaps between the data available and the data needed has followed a priority scale established with SP2, SP3, SP4, SP5, SP6, CDP1 and CDP2. The modelling work and the coordination in particular with SP6 and CDP2, ensures that the data generated are critical for the development of specific HBP models/simulations. This plan is based on the document *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020* (Version 3.0, 26 July 2016). The first version of the SP1 data management plan (SGA1) was delivered in M20 and was updated and submitted by M24 (as scheduled); the current version is an updated version as requested in the SGA1 Final Review that includes further information on the role of the acquired data regarding in which SPs the data are/were going to be used), as well as other required updates. The plan is being implemented in the SGA2 and will continue throughout the project.

Keywords:	data management plan, datasets, FAIR data, models, reports, software

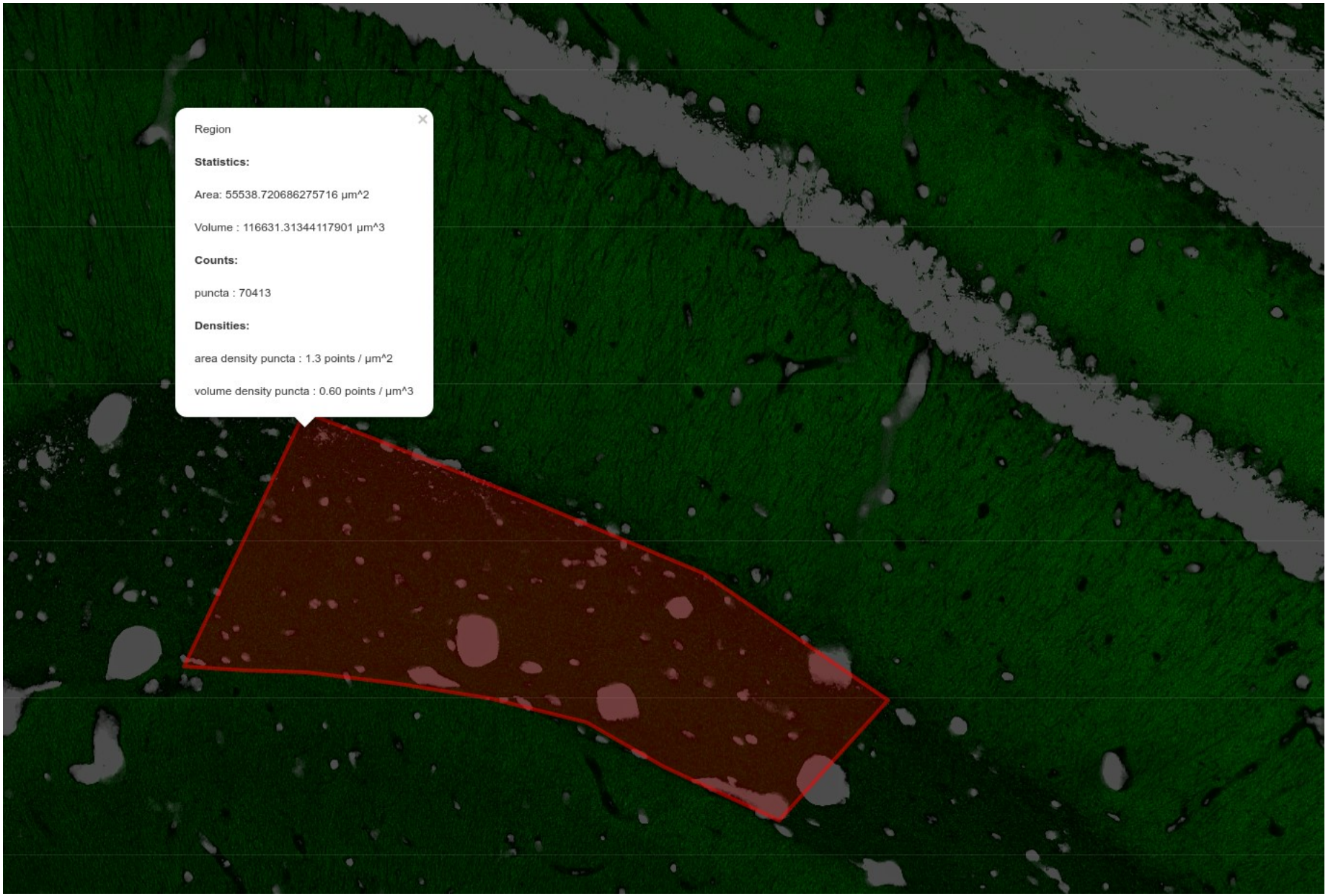

Figure 1: Graphical User Interface Detail from Multimap (SP1 Mouse Brain Organisation)

Targeted users/readers	HBP SPs
Contributing Work-Package(s):	SGA1 WPs 1.1, 1.2, 1.3, 1.4, 1.5
Initially Planned Delivery Date:	SGA1 M24 / 31 Mar 2018 (Date for submission to EC, as set out in DoA) 31 July 2018 8Date for resubmission set out in the SGA1 Review Report)

Authors:	Pilar F. ROMERO, UPM (P68); Javier DEFELIPE (P68)
Compiling Editors:	Pilar F. ROMERO, UPM (P68);
Contributors:	Rafael LUJAN, UCLM (P65) Ryuichi SHIGEMOTO, IST (P31) Simon BERNECHE, SIB (114) Antonino CATTANEO, SNS (P116) Enrico QUERUBINI, EBRI (P115) Michele MIGLIORE, CNR (P12) Douglas ARMSTRONG, UEDIN (P62) Javier DEFELIPE, UPM (P68) Huib MANSVELDER, VU (P113) Sten GRILLNER, KI (P37) Egidio D'ANGELO, UNIPV (P70) Tamás FREUND, Szabolcs KALI, IEM HAS (P30) Zoltan KISVARDAY, UoD (P15) Francisco CLASCA, UAM (P64) Ángel MERCHÁN, UPM (P68) Francesco PAVONE, LENS (P40) Leonardo SACCONI, INO, CNR (P12) Bruno WEBER, UZH (P75) Concha BIELZA, UPM (P68) Luis PASTOR, URJC (P69)
SciTechCoord Review:	EPFL (P1) Jeff MULLER
Editorial Review:	EPFL (P1): Guy WILLIS, Annemieke MICHELS

Table of Contents

Table of Figures

Table of Tables

Introduction

This plan is based on the document *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020* (Version 3.0, 26 July 2016). Each entry includes, among other information, the planned use of data in models; estimated date of delivery; timeline for curation and annotation; and a specification of the target quantity and quality. The plan also includes the mechanism and the periodicity of how the list of deliveries was updated during SGA1 in the Data Catalogue (see below).

A first version of this plan was produced in M18 (HBP-SGA1-SP1DMP-M18-v4) and updated in M23 (HBP-SGA1-SP1DMP-M23-v3.4). In M28, July 2018, the report has been improved and updated via this document as requested. Finally, this plan is in line with and linked to the current version of the HBP Data Management Plan (SGA1 Deliverable D11.3.2).

This plan includes an explicit list of data deliveries by SP1 in SGA1 and is composed of two sections: (i) general information common to all datasets, and (ii) specific information for each dataset (Annex 1: list of datasets)

Detailed information on the role and take up of the data acquired can be found in Annex 1 for each dataset listed as well as in the SP1 Data Catalogue.

This plan is complemented with the SP1 SGA1 Data Catalogue (see document SP1 Data Catalogue Addendum to SGA1 PPR M01-M24) currently including the datasets, as well as tools, reports and models, generated up to M23[1]*. The updated version of the Data Catalogue was released by M24 as an attachment to the SGA1 M1-M24 Project Periodic Report.

SP1 Data Management Plan

This data management plan (DMP) is for SP1 in the SGA1 phase of the HBP. It has been prepared following the guidance provided by the document H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 (Version 3.0, 26 July 2016). This implies that the DMP is a living document, and will be updated periodically. The first version produced was HBP-SGA1-SP1DMP-M18-v1.4, an updated version was submitted by M24 (HBP-SGA1-SP1DMP-M23-v3.4), and the current version is HBP-SGA1-SP1DMP-M23-v4.4.

The DMP is complemented with the Data Catalogue (available at https://collab.humanbrainproject.eu/#/collab/5972/nav/105733 (Storage/SGA1/SP1 Data Catalogue), and also added to the Switch drive)).

As a number of the elements of the DMP are common to the whole HBP, for the sake of brevity, in places this DMP refers to the HBP Data Management Plan (SGA1 Deliverable D11.3.2, delivered on 16 June 2017), which is also a living document and a new version has been submitted in SGA1 M24. This document starts with information that is shared between all SP1 datasets, and then lists datasets produced by SP1, with fields describing information specific to those datasets.

Table 1: Component Details for SP1 Data Management Plan

	Main Meta Data	Comment/title
Component	1015	D1.5.3: Detailed plan of data usage and the impact of generated data on models
Component Type	Report	
Contact	T1.5.1 (DEFELIPE, Javier, F. ROMERO, Pilar)	

[1] M23: target month to submit the deliverables to the PCO according to its instructions

Latest Release	1/03/2018	
TRL	NA	
Location	https://collab.humanbrainproject.eu/#/collab/5972/nav/105733	Also shared with NIP
Maintenance	TBD	
Curation Status	NA	
Validation - QC	Yes	SP1 members, SP5, and SP11
Validation – Already existing users	Yes	SP1, SP2, SP4, SP5, SP6, SP10
Validation – Use in publications	No	
Access privacy	NA	
Access sharing	NA	
Access licence	NA	
URL to access component	https://collab.humanbrainproject.eu/#/collab/5972/nav/105733	
URL to component documentation	https://collab.humanbrainproject.eu/#/collab/5972/nav/105733	
URL to component usage documentation	https://collab.humanbrainproject.eu/#/collab/5972/nav/105733	
URL to dissemination material highlighting component		

HBP-SGA1-SP1DMP-M23-v4.4

Specifically, the SP1-DMP is composed of 7 main sections as follows:

1) Data summary: this section outlines the purpose of (i) data collection/generation and (ii) the software and models to be developed. The relation with the SP1 SGA1 objectives is also included. In addition, the information regarding type, formats, re-use, origin, expected size, and utility of the SP1 datasets is included.
2) FAIR data: this section outlines how to make data findable, accessible, interoperable, and reusable (FAIR) according to the document *H2020 Programme Guidelines* mentioned above.
3) Allocation of resources: this section explains the allocation of resources addressing estimated costs to make SP1 data FAIR, responsibilities of this DMP and the potential value of long term preservation.
4) Data security: data recovery as well as secure storage and transfer of data is described.
5) Ethical aspects: in this section, the status of the ethical issues is clarified referring to the DoA.
6) Other: other relevant information on the SP1 datasets
7) Annex 1: List of Datasets

Although this plan is mainly focused on data, the information concerning the IT tools and statistical models that have been — or are currently being — developed in the SGA1 for SP1 is also included in the above items.

The software packages developed in SP1 are also included in the HBP Software Catalogue located at the HBP Collaboratory: https://collab.humanbrainproject.eu/#/collab/19/nav/2108

1. Data summary

1.1 Purpose of data collection/generation

In general, SP1 data is collected to support modelling and analysis elsewhere in the HBP, particularly SP6 and SP4, and beyond. SP1 data are critical for internal HBP interactions. SP1 generates data for subcellular modelling, for neuronal model reconstruction and validation, for microcircuit model reconstruction and validation, for whole-brain model reconstruction and validation, and for databasing. We generate data of high granularity that are required for the simulations in SP6 and CDP2 and other SPs and CDPs, and do not exist elsewhere.

Data on the detailed morphology of individual neurons is needed to construct realistic models of neurons, which also requires information on ion channels and receptors expressed in the somatodendritic membrane. Electrophysiological data in terms of membrane properties on the same cells are needed to construct cell models that are as close as possible to their biological counterparts. Data on distributions of synapse types on dendrites and soma, both on individual neurons and throughout the brain is needed, to build realistic models of neuron types and circuit models. The focus in SP1 is to investigate the microcircuits, in four major areas of the brain, the hippocampus, neocortex, basal ganglia and cerebellum. SP1 provides strategic missing pieces required for the modelling efforts in SP6. Data is also collected at the level of the whole brain to link the subcircuit together starting from SGA1 but even more so in subsequent phases on the project. Section 1.7 and the detailed tables below provide more specific information about which datasets are used by whom. Proteomic and molecular interaction data are collected both by experiment and by curating public databases and the literature.

SP1 also develops software packages. The purposes of these packages are (i) to analyse morphological data, where domain knowledge is especially useful for developing software packages; (ii) to analyse protein-protein network structure, where the methods of clustering require domain knowledge; and (iii) to provide a framework to formalise protein-protein interactions in such a way that they may be incorporated easily into subcellular molecular models.

1.2 Relationship to SP1's Objectives for SGA1

All the datasets that SP1 expects to collect, as well as the software and models it has developed in SGA1 are directly linked to SP1's objectives in this project phase.

Table 2: SP1 Objectives for SGA1

SP1	SGA1 Objective
1a	Maps of the vasculature of the whole mouse brain.
1b	Whole-brain maps of different cellular types based on gene expression.
1c	Microcircuitry analysis, proteins and receptor distributions, fibre architecture in a specific large area brain region.
1d	Maps of cellular distributions, long-range axonal projections, and synaptic proteins; reconstructed morphologies of major neuron types.
1e	Whole-brain activation maps related to selected behaviours

1f	Spatial organisation principles in brain activation
1g	Functional maps of cortical activity during learning of the motor task after stroke during learning in the robotic platform

The specific relationship of each dataset, software, report or model with the SP1 objectives is displayed the description of each dataset in Annex 1: List of datasets. Some datasets also relate to objectives linked to other HBP Subprojects.

1.3 Types and formats of data generated/collected

We are generating 12 broad types of data, including software and models (see Table 3).

Table 3: Types and formats of data generated

Type	Formats
Image	tiff, jpg, png
Image stack	tiff, mp4
Electrophysiological recording	abf, brw, ,rsh, rsm, rsd, dha, tbk, spd
Morphological reconstruction of neuron	asc, xml, dat
Whole-cell current clamp recording	tsv
Point-clouds of coordinates	txt
Segmentation data	seg, xls
Parameters	csv
Molecular binding and state transition rules	ka
Bioinformatic meta-analysis	tsv, csv
Model	ipynb, py, ka
Software	py, cpp, java

Each type of data can be stored in one or more formats. In Section 2.2.3, we list the software corresponding to each format.

1.4 Data Re-use

Most of datasets are not reusing data, reflecting the focus of the subproject on data collection. Datasets in which data are being re-used include software datasets, which are developing existing code, and molecular curation datasets.

Table 4: Data re-use

Reused	Freq
Maybe	8
No	31
Yes	7

1.5 Origin of the data

The list of datasets (Annex 1) contains information for each dataset, including the data class (e.g. cellular, molecular or electrophysiological), text describing the specimen information and the brain region and subregion.

An analysis of the data class is shown below.

Table 5: Data class

Data class	Number of dataset
Cellular	16
Circuits	1
Electrophysiological	6
Model	1
Molecular	12
Other	3
Software	1
Software and models	1
Subcellular	3
Whole-brain activation maps	1
Whole-brain cell distributions	1

The largest class is "cellular", reflecting a number of tasks that are reconstructing morphologies.

The predominant experimental specimen is the C57/B6J mouse, at a variety of ages, though human postmortem tissue is also being collected in some datasets, and one dataset is collecting data from Wistar rats. Some datasets are partially derived from data in the literature, and the software datasets do not have any origin in this sense.

An initial analysis of the brain regions mentioned is shown below. Some datasets include information from more than one brain region, so the total number is greater than the number of datasets.

Table 6: Brain regions

Brain region	Number
basal ganglia	1
cerebellum	2
Cerebral cortex	1
cortex	3
forebrain	2
hippocampal formation	5
hippocampus	15
hypothalamus	1
motor cortex	1
n/a	1
neocortex	4
not applicable	1
somatosensory cortex	10
temporal cortex	1
thalamus	1
ventral midbrain	1
visual cortex	2
whole brain	2

Once data collection starts, these metadata will be collected as part of the data curation process (see Section 2.0).

1.6 Expected size of the data:

The expected size of the data is about 53 TB for 46 datasets. The expected size per each dataset is displayed below.

Table 7: Size of the data

Task	Dataset Title	Owner	Size (TB)
T1.1.1	Developing the integrated FIB/SEM and SDS-FRL immunoelectron microscopy technique.	Rafael Luján & Riuichi Shigemoto	2.0e-02
T1.1.1	Nanoscale measurements of distribution of individual receptors and ion channels in cortical neurons.	Rafael Luján & Riuichi Shigemoto.	2.0e-02
T1.1.2	Generation of new intrabodies / antibody fragments	Antonino Cattaneo, Giovanni Meli	1.0e-06
T1.1.2	IACT antibody fragments for imaging	Antonino Cattaneo, Giovanni Meli	1.0e-03
T1.1.3	Association (co-clustering) of receptors and their effector ion channels in different neuronal compartments.	Rafael Luján & Riuichi Shigemoto	2.0e-02
T1.1.3	GluA2, GluA3 and GluN1 in CA1	Ryuichi Shigemoto	1.0e-02
T1.1.4	Electrophysiological data-hippocampus	Enrico Cherubini	1.0e-03
T1.1.5	K channel kinetic and activity in a model neuron	Simon Bernèche	1.0e-04
T1.1.6	Curated list of synaptic protein-protein interactions	Oksana Sorokina	1.0e-03
T1.1.6	Computational, dynamic model of LTP and LTD in a wild type Schaffer Collateral synapse	David Sterratt	1.0e-03
T1.1.6	KappaNEURON software package	David Sterratt	1.0e-03
T1.1.6	A mapping of computational models of synapses to proteins	David Sterratt	1.0e-03
T1.1.7	Subcellular proteomics dataset - hippocampus	Antonino Cattaneo	2.0e-03
T1.1.7	Synaptic Plasticity dataset - hippocampus	Enrico Cherubini	1.0e-01
T1.1.7	Extending coverage of published data	Douglas Armstrong, Antonino Cattaneo	1.0e-03
T1.1.7	Genetic mapping to single cell profiles	Douglas Armstrong, Antonino Cattaneo	1.0e-03
T1.1.7	Integration of functional data into synapse models	Douglas Armstrong, Antonino Cattaneo	1.0e-03

Task	Dataset Title	Owner	Size (TB)
T1.2.1	3D reconstructions mouse hippocampus	Javier DeFelipe	5.0e-03
T1.2.1	3D reconstructions rat hippocampus	Javier DeFelipe	5.0e-03
T1.2.1	3D reconstructions pyramidal neurons human cortex	Javier DeFelipe	5.0e-03
T1.2.1	3D reconstructions human hippocampus	Javier DeFelipe	5.0e-03
T1.2.1	3D reconstructions pyramidal neurons mouse cortex	Javier DeFelipe	5.0e-03
T1.2.2	human neurons with matching Ephys	Mansvelder	0.0e+00
T1.2.3	The striatal microcircuit	Sten Grillner	5.0e-03
T1.2.4	Electrophysiological data cerebellum	Egidio D'Angelo	1.0e-03
T1.2.5	Morphological database of major cell types of the mouse hippocampus	Szabolcs Káli	1.0e-01
T1.2.5	Electrophysiological database of major cell types of the mouse hippocampus	Szabolcs Káli	1.0e-03
T1.2.5	Morphological reconstructions of mouse hippocampal neurons filled in vivo	Szabolcs Káli	1.0e-02
T1.2.5	Physiological characterisation of mouse hippocampal neurons recorded in vivo	Szabolcs Káli	5.0e-03
T1.2.6	Database of synaptic physiological properties in the mouse hippocampus	Szabolcs Káli	1.0e-04
T1.2.7	GABAergic neuron subtypes	Zoltán Kisvárday	1.0e+00
T1.2.8	3D digital Reconstructions of individual thalamocortical neurons	Franciscp Clasca (UAM)	1.0e-02
T1.2.9	Densities and 3D distributions of synapses using FIB/SEM imaging in the human neocortex (Temporal cortex, T2)	MERCHÁN PÉREZ, Ángel	5.0e-03
T1.2.9	Densities and 3D distributions of synapses in the human hippocampus	MERCHÁN PÉREZ, Ángel	5.0e-03
T1.2.9	Densities and 3D distributions of synapses in the mouse neocortex	Angel Merchan Perez	5.0e-03
T1.2.9	Densities and 3D distributions of synapses in the mouse hippocampus (CA1)	Angel Merchan Perez	5.0e-03
T1.2.9	Immunocytochemical detection of excitatory and inhibitory terminals in the mouse somatosensory cortex by confocal microscopy	Alberto Muñoz	5.0e-03
T1.2.9	Immunocytochemical detection of excitatory and inhibitory terminals in the mouse hippocampus (CA1) by confocal microscopy	Alberto Muñoz	5.0e-03
T1.3.1	Whole brain interneuron distributions	Ludovico Silvestri	4.1e+01
T1.3.2	Wide-field imaging of cortical activity during motor learning	Anna Letizia Allegra Mascaro	5.0e-04

Task	Dataset Title	Owner	Size (TB)
T1.3.2	Wide-field imaging of cortical activity one month after stroke and rehabilitation	Anna Letizia Allegra Mascaro	1.1e+01
T1.3.2	Wide-field imaging of cortical activity one month after stroke	Anna Letizia Allegra Mascaro	5.0e-04
T1.3.4	Whole brain maps of resting state brain activation	Ludovico Silvestri	1.1e+01
T1.3.5	3D image of the vascular system of the mouse brain	Velizar Efremov	1.3e-02
T1.3.5	Model of intravascular and tissue partial pressure of oxygen	Velizar Efremov	0.0e+00
T1.3.5	3D reconstruction of the vascular system of the mouse brain	Velizar Efremov	5.0e-04
T1.4.1	Analysis of micro-anatomical data	Concha Bielza	0.0e+00
T1.4.2	Software PyramidalExplorer 1.2 for early exploratory analysis techniques for morphological data.	Universidad Rey Juan Carlos	5.0e-02

1.7 Data utility

SP1, in agreement with the modelling pipeline of the HBP, will focus on four major brain circuits: neocortex (including the thalamocortical system), hippocampus, basal ganglia and cerebellum. The work plan will focus on fundamental questions, coordinated at the HBP level, on structural organisation, neuronal activity, microcircuit dynamics, synaptic plasticity and neuromodulation required to fuel and complement modelling and theory. The data will serve for both model reconstruction and validation in virtuous feed-back cycles between simulations and experimental recordings (data generated models and simulations instruct hypothesis-driven data sampling). The data will also be used to obtain high-quality integrative maps and circuits of the mouse brain at the functional and anatomical levels integrating SP5 databasing and brain atlasing with molecular, morphological and functional data. An analysis that extends from gene to behaviour needs to be based on animal data combined with human experimentation (see Figure 2 and Figure 3).

The particular role of SP1 is to supply strategic information, not yet available, on the molecular, cellular and synaptic level that is needed in particular for the simulations/models of SP6 and SP4 and also other SPs. The simulations of SP6 are of course based predominantly on experimental data from the scientific literature or through e.g. Allen Brain. To supply still missing, specific detailed data required for both the subcellular and the large microcircuit is the task of SP1. It is important to understand that the granularity of the information required for the detailed SP6 simulation cannot be met by general databases such as the Allen Brain, although they contribute importantly to the overall knowledge of these circuits.

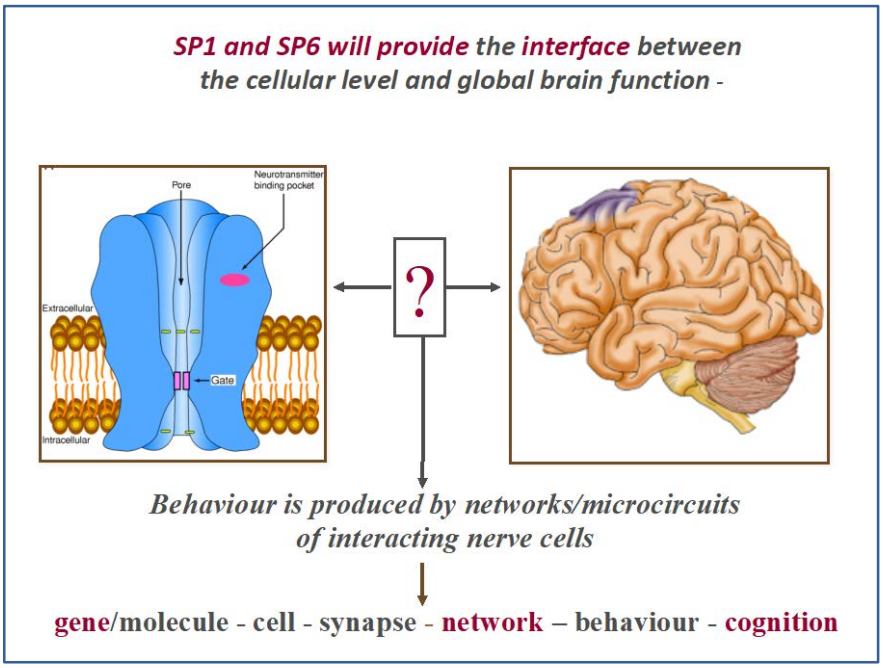

Figure 2: From Gene to Behaviour

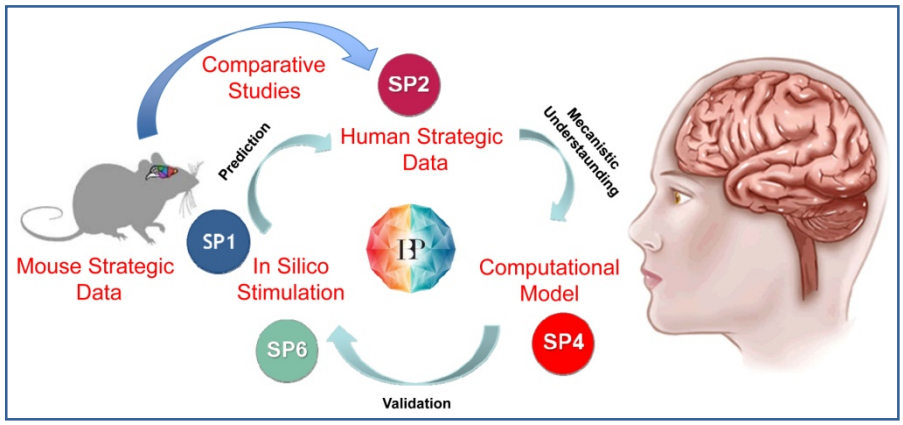

Figure 3: From SP1 data to HBP models and theories on the human brain

Each WP and task in SP1 maps on to the needs of specific tasks in SP6 or SP4 or other SPs as specified in detail in the ***"data catalogue"*** available at https://collab.humanbrainproject.eu/#/collab/5972/nav/105733 (Storage/SGA1/SP1 Data Catalogue), and also added to the Switch drive.

- WP 1.1 "*Subcellular and molecular data"* provides strategic information that primarily is critical for both WP6.1. Subcellular and Molecular modelling tasks 6.1.1/1.2/1.3 and WP6.2 (6.2.1 – 6.2.5) and also T 4.1.1. WP1.2
- WP 1.2 *Cell and Microcircuitry: neocortex, hippocampus, basal ganglia and cerebellum*. This WP provides strategic data onto WP 6.2 and tasks 6.2.2 – 6.2.5 which provides models for the four microcircuits, and also to CDP2
- WP1.3 *Whole brain*, this WP delivers data to T6.2.3, T6.2.6, T 4.1.4, T 4.4.5, T 4.4, T 4.5.1, T 4.5.2, T10.1.3
- WP 1.4 *"Microanatomical, structural and functional integration of data in brain circuits"* provides input to in particular T6.2.2, T6.2.3, T6.2.7 and CDP2
- WP1.5 "*Comparative study of cells and microcircuits in the rodent and human brain"* have supported T.4.1, T.6.2.2, T6.2.4

SP1 data collection, organisation and utility are in line with the data objectives for HBP displayed in the HBP DMP, the main purposes of which can be summarised as follows:

- *Collection* - collect data experimentally to satisfy a scientific use case.
- *Organisation* - aggregation and curation for the purpose of making data more readily
- Reusable
- *Model building* –SP1 data are required for developing detailed cellular models of individual neuron types to be used in e.g. microcircuit simulation in SP6.
- Model validation – SP1 data is used to validate existing models.
- Model exploitation - brain-inspired models are used for a variety of use cases, including but not limited to:
  - validating data collection
  - performing *in silico* experiments
  - configuring neuromorphic hardware
  - controlling virtual and real robots
- Tool development - data used for the development of software tools.

The specific utility of the SP1 data is displayed in each of the dataset description in Annex 1: List of datasets). We have indicated the utility to HBP partners and to the wider world.

2. FAIR data

2.1 Data Curation

SP1 Partners and SP5 are jointly responsible for data curation. The curation process can be summarised as follows:

1) SP1 Partners submit data and metadata to the HBP Collaboratory (for smaller amounts of data) or the CSCS (SP7) data containers (for larger amounts).
2) Basic metadata are organised and stored in the Data Workbench.
3) The SP5 Tier 1 curation team validates basic metadata entries for completeness.
4) Once metadata curation has been finished for a given dataset, SP5 will know whether the data are suitable for more specialized curation (one of the Tier 2 types).
   a) SP5 Tier 2A curation teams validate spatial metadata for accuracy and completeness. Spatial coordinates will be assigned in Mouse Brain Atlas space to data, where possible.
   b) The SP5 Tier 2B curation team validates neural activity data for accuracy and completeness.
5) After validation, all metadata are organised and stored in the Knowledge Graph
6) Data in the Knowledge Graph are available for query, viewing and analysis through platform services.
7) Spatial data are used for displaying data in atlas space.

SP5 estimates that, depending on the demand for and the depth of curation, Tier 1, curation will take 1-2 weeks at present and Tier 2 will take between 2 days and 3 weeks. SP5 estimates the whole curation process to take between 1 and 5 weeks with the current workflow.

In more detail:

The curation team will be led by SP5. When everyone is informed by SP1, SP5 will first have a one-to-one call with the contact person for the group/lab/unit to clarify. SP5 will then ask for a video-conference or Skype call to walk through the data that have been uploaded in the Collaboratory or container. Each group decides how many people will participate. From the curation team, there will be three primary participants in the curation session. Others may also participate - but these three curation team members will be responsible for leading the curation session. SP5 will send out information on any preparation needed, including the testing of an online application for entering metadata. The curation team will ensure that this application meets the needs of all data categories and hierarchies. When metadata are assigned, the complete list will be delivered back to the neuroscience groups for a final check. It will be SP5's responsibility to make the metadata searchable as soon as possible.

Data access: the process outlined above does not result directly in any sharing of data. Before sharing, the groups will be given the opportunity to choose a sharing license.

Embargo: after a license has been chosen, there may still be a need for an embargo period for some data. Details will be worked out taking into account requirements from the European Commission.

2.2 Making data findable, including provisions for metadata

2.2.1 Discoverability of data (metadata provision)

We will supply metadata to SP5 who will then enter it into the Knowledge Graph, to allow us to tag datasets with metadata. This is part of the Tier 1 curation process. http://interlex.org/.

2.2.2 Identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

When data is made public, we will endeavour to provide a DOI for it. The HMP DMP states: "It is a goal to enable data sharing with publicly released datasets given immutable, unique identifiers. However, the details of this capability will vary from storage service to storage service." This item will be discussed with SP5 and SP11.

2.2.3 Naming conventions used

As per the HBP SP11 DMP (SGA1 D11.3.2), file naming conventions are generally not required for data handled by the HBP. Because of the diverse nature of the data contained in SP1, it would be difficult to write a convention. However, it is strongly recommended that, where naming conventions are important for data reuse or interpretation, they should be adequately documented in a file named README, provided with the data.

Any naming conventions will also be documented in the Tier 1 metadata during the curation process.

2.2.4 Approach used for search keywords

The searchability of SP1 data will be assured by following the general approach described in the HBP SP11 DMP (Section 4.1.1) that is as follows:

Data can be discovered through the Web-based user interfaces provided by the SP5 NIP. These interfaces will allow discovery by browsing or by search. Generally speaking, the Access Control Lists of the Collaboratory will be respected to allow user control of privacy settings. In the places where there are caveats to this behaviour, as may be the case in some FENIX data repositories, this will be clearly indicated in the user service documentation. Data will be searchable using SP5 search tools integrated into the HBP Collaboratory. This search will honour Collab membership ACLs in the return of search results. The search will be available on the following content fields:

- Collab metadata (name and description)
- Collab app content (for Collaboratory provided apps, where the app requests indexing)
- Platform app content (where enabled by Platform apps)
- Software Catalogue entries (name, description and tag fields)
- Dataset metadata as defined the by the HBP minimum metadata specifications developed in SP5

2.2.5 Approach for clear versioning

The approach for versioning needs to be discussed with SP5.

2.2.6 Standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

As shown in the Data Summary, we are dealing with diverse types and classes of data. The HBP DMP proposes having a "minimal" or "basic" metadata scheme, common to all datasets, which will be supplemented by domain-specific metadata. The Tier 1 curation undertaken in collaboration with SP5 will collect the basic metadata. The Tier 2 curation will collect domain-specific metadata, such as spatial location.

We have already collected the basic metadata as part of this DMP. We have yet to agree with SP5 which standards to use for domain-specific metadata. As expanded on in Section 2.3, where possible we will use existing ontologies or lexicons to specify items clearly.

2.3 Making data openly accessible:

2.3.1 Data that will be made openly available

The SP1 DMP will follow the statement concerning the European policy on Open Data contained in the HBP FPA as a guiding principle of the HBP's Data Governance Policy and will be respected in all policy and data repository service implementations, as confirmed in the HBP SP11 DMP (SGA1 D11.3.2 Section 4.2.1).

According to the HBP SP11 DMP, in addition to inaccessibility of data required by compliance with European and National legislation, certain classes of data may remain inaccessible for various other reasons. Animal data produced in the HBP may remain temporarily inaccessible to users outside the HBP if:

- Data is a critical competitive dependency for a pending publication.

Human data produced in the HBP may remain temporarily inaccessible to users outside the HBP if:

- Data is a critical competitive dependency for a pending publication.

Human data produced in the HBP may remain permanently inaccessible to users outside the HBP if:

- Data was not collected with sufficient releases to allow sharing outside the HBP.
- Data might be easily re-identifiable when combined with current technology and other readily available datasets.

(See Section 4.2.1 in the HBP SP11 DMP for further information)

2.3.2 How the data will be made available

As for the item above, the SP1 DMP will follow the HBP SP11 DMP with regard to accessing data (see HBP DMP, Section 4.2.2 for further information).

2.3.3 Methods or software tools needed to access the data

Methods and software to access the data as well as the relevant documentation as describe in the following table:

Table 8: Methods and Software

Extension	Format Name	Documentation URL	Vendor	Software to open format
tiff	Tagged Image File Format	https://en.wikipedia.org/wiki/Tagged_Image_File_Format	Open standard	ImageJ, gimp, many open source libraries
txt	Text file	http://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap03.html#tag_03_397	Open standard	Many text editors, genral purpose packages (e.g. Python)
jpeg	Joint Photographic Experts Group	https://en.wikipedia.org/wiki/JPEG	Open standard	ImageJ, gimp, many open source libraries
mp4	MPEG-4 Part 14	https://en.wikipedia.org/wiki/MPEG-4_Part_14	Open standard	Many libraries, and packages, e.g. libxml, R XML package
mat	Matlab binary format	https://www.mathworks.com/help/pdf_doc/matlab/matfile_format.pdf	Mathworks	Matlab, R, Octave
xls	Microsoft Excel worksheet sheet (97–2003)	https://msdn.microsoft.com/en-us/library/gg134032.aspx	Microsoft	Microsoft Excel, Libreoffice, R xlsconnect package
dat	Neurolucida binary format	http://www.mbfbioscience.com/help/si11/Content/About/FileFormats.htm	MBF Bioscience	NeuroConstruct (http://www.neuroconstruct.org/docs/import.html#Neurolucida) can import ASCII (*.asc) format V3 files
abf	Axon Binary Format	https://moleculardevices.app.box.com/s/iisgk109swvcrwtmy3p13dn3r0vyvasr	Molecular Devices	Abfload (Matlab script; https://uk.mathworks.com/matlabcentral/fileexchange/6190-abfload)
brw	BrainWave	http://www.3brain.com/websites/3brain/downloads/BioCAM-X_userguide.pdf	3Brain	http://www.3brain.com/downloads.html#software-and-utilities

Extension	Format Name	Documentation URL	Vendor	Software to open format
mod	nmodl	https://www.neuron.yale.edu/neuron/static/docs/help/neuron/nmodl/nmodl.html	Duke, Yale, BBP	https://www.neuron.yale.edu/neuron/static/docs/help/neuron/nmodl/nmodl.html#ModelDescriptionLanguage
ipynb	IPython Notebook	https://en.wikipedia.org/wiki/IPython	Open source software	Jupyter notebook
rsh	BrainVision	http://www.scimedia.com/fis/support/download/ultima/	SciMedia	BV_Ana (http://www.scimedia.com/fis/support/download/ultima/)
rsm	BrainVision	http://www.scimedia.com/fis/support/download/ultima/	SciMedia	BV_Ana (http://www.scimedia.com/fis/support/download/ultima/)
rsd	BrainVision	http://www.scimedia.com/fis/support/download/ultima/	SciMedia	BV_Ana (http://www.scimedia.com/fis/support/download/ultima/)
dha	BrainVision or Matlab	http://www.scimedia.com/fis/support/download/micam01/	SciMedia	BV_Ana (http://www.scimedia.com/fis/support/download/micam01/)
tbk	Tucker-Davis Technologies	http://www.tdt.com/files/manuals/OpenEx_User_Supplement_Syn.pdf	Tucker-Davis	OpenEx Software (http://www.tdt.com/openex.html)
spd	SpikeTrain	http://www.neurasmus.com/spiketrain/SpikeTrainUserGuideV1_08.pdf	Neurasmus	SpikeTrain (http://www.neurasmus.com/spiketrain.php)
png	Portable Network Graphic	https://en.wikipedia.org/wiki/Portable_Network_Graphics	Open standard	ImageJ, gimp, many open source libraries
xml	Extensible markup language	https://en.wikipedia.org/wiki/XML	Open standard	Many libraries, and packages, e.g. libxml, R XML package
docx	Microsoft Office XML document	https://en.wikipedia.org/wiki/Microsoft_Office_XML_formats	Microsoft	Microsoft Word, Libreoffice Writer
seg	Espina	http://cajalbbp.cesvima.upm.es/espina/	Universidad Rey Juan Carlos	Espina (http://cajalbbp.cesvima.upm.es/espina/)
csv	ASCII text as comma-separated values	https://en.wikipedia.org/wiki/Comma-separated_values	Open standard	Microsoft Excel, Libreoffice Calc, many general purpose packages (e.g. Python, R, Matlab)
tsv	Tab-Separated Values	https://en.wikipedia.org/wiki/Tab-separated_values	Open standard	Microsoft Excel, Libreoffice Calc, many general purpose packages (e.g. Python, R, Matlab)

Extension	Format Name	Documentation URL	Vendor	Software to open format
ka	Kappa file (SpatialKappa dialect)	https://github.com/lptolik/SpatialKappa/raw/master/docs/manual/SpatialKappaManual-v2.1.0.pdf	Open standard	SpatialKappa (https://github.com/lptolik/SpatialKappa), KappaNEURON (https://github.com/davidcsterratt/KappaNEURON)
gml	Graph Modelling Language	http://www.fim.uni-passau.de/index.php?id=17297&L=1	Open standard	Gephi (https://gephi.org/)

This needs to be documented in the Tier 1 metadata "during the curation process". Further information will be provided in the next version of this plan.

2.3.4 Data and associated metadata location, documentation and code needed

SP1 data will initially exist on laboratory servers. As described in Section 2.1, it will then be moved either to the HBP Collaboratory (for smaller datasets, in the GB size range) or the containers of the CSCS servers (SP7) following the approach set up. In addition, the data may be mirrored on subject-area or institutional repositories.

The metadata will exist in the HBP Knowledge Graph, having been curated by SP5, as described in Section 2.1.

Documentation will be included with the datasets, usually in the form of a README file.

Code in the form of software packages will be deposited in source control repositories, such as GitHub or Launchpad. Code in the form of analysis or model scripts will be deposited with the associated dataset or model. Documentation for software packages should be included with the code.

The SP1 DMP data list will provide concrete data status updates and provide downloadable links to the location of data, as well as the documentation for each respective data set, as available.

2.3.5 How access will be provided in case there are restrictions

Any restrictions on access to SP1 data need to be documented in the Tier 1 metadata "during the curation process". The classes of restriction specified in the SP1 DMP are as follows:

- privacy (e.g. Human data)
- very large file size (e.g. high-resolution images)
- a publication embargo
- other

More details will be provided in the next version of this DMP.

2.4 Making data interoperable

2.4.1 Assess the interoperability of your data

There is a common approach for Tier 1 and 2 metadata, based on the formats described above.

The Tier 2a curation, which anchors data in the Brain Atlas, will allow data interoperability.

Regarding vocabularies, we will work with the HBP Ontology Definition Team (Section 4.3.2 of HBP DMP) to determine which existing ontologies to use or which new ones are needed. In the meantime, we will use Interlex (http://interlex.org) to identify Brain Regions, Cell Types and Techniques.

2.4.2 Standard vocabulary used for all data types present in a given data set, to allow inter-disciplinary interoperability

Much of our data is images or numerical (e.g. electrophysiology recordings) and therefore standard vocabularies are probably not so relevant in our case.

To identify genes in our molecular data, we will use standard gene names, as defined by HGNC (http://www.genenames.org) and the Mouse Genome Informatics database (http://www.informatics.jax.org/), and federated by the Entrez Global Query Cross-Database Search System (https://www.ncbi.nlm.nih.gov/). We are exploring which are the best vocabularies to use for molecules. Work in SGA1 to find a suitable ontology to describe molecules used in models and to map them to genes suggests that such an ontology does not exist. Consistent use of gene identifiers allows us to make use of resources such as OMIM (diseases) and gene ontologies. We will provide mappings from Human to Mouse genes.

2.5 Increase data re-use (through clarifying licenses)

2.5.1 How the data will be licenced to permit the widest reuse possible

The HBP SP11 DMP states that "...the HBP strongly encourages public release of any and all reusable experimental data. However, the understanding of the importance and career value for individual researchers is expected to grow slowly outside consortia such as the HBP. As a result, data licences will be chosen by the data providers (both inside and outside the HBP) primarily, selecting from a subset of Creative Commons licenses." It goes on to suggest seven Creative Commons licences. We have suggested these to SP1 Partners, along with an "another" option.

Table 9: Licensing of data

Data licence type	Freq
All rights reserved, Copyright	2
Attribution NonCommercial ShareAlike 4.0 International	4
BY - Attribution alone	2
BY-NC - Attribution + Non-commercial	19
BY-NC- ND - Attribution + Non-commercial + No Derivatives	5
BY-NC-SA - Attribution + Non-commercial + ShareAlike	5
BY-SA - Attribution + ShareAlike	2
CC0 - Freeing content globally without restrictions	1
GPL3	1
MIT	1
We need to match BIOGRID, IntAct and DIP terms & licences to CC licenses	4

A total of 40 datasets are to be licenced under one of these CC licences, most of them under BY-NC, 1 software package is released under GPL3 and the remaining under other licences. The

ramifications of choosing a particular licence took some time to work though, so we planned to have a discussion on this with SP11, to be in line with the general HBP DMP.

2.5.2 When the data will be made available for re-use

See "RELEASES" section of each dataset in Annex 1 for expected availability of data and embargo information.

2.5.3 Re-use of the data produced and/or used in the project by third parties, in particular, after the end of the project

Data will be re-useable by third parties after the Project, subject to privacy issues and reasonable publication embargoes. Each dataset indicates the privacy class. There may be issues concerning reuse of Human Research data, but we have not yet explored the issues in the datasets marked as being Human Research.

Table 10: Privacy class

Privacy class	Number of datasets
Animal Research	32
Human Research	6
No privacy constraints	7

2.5.4 Data quality assurance processes

Each lab has its own quality assurance mechanisms (see per-dataset notes in Annex 1). In general, SP1 data are also of interest in its own right, and will therefore be presented in journal submissions and subject to the usual scientific peer review process. The curation process will encourage further checking of data by producers. Sharing data with HBP Partners also provides a form of peer review of the data.

2.5.5 Length of time for which the data will remain re-usable

We understand that reusability involves a commitment to make sure that software is supported to use the files. Most of the formats that we use are open standards (e.g. tiff, csv), which will be readable for the foreseeable future. A number of the proprietary formats have open-source readers available, hosted on sites that are likely to remain live for the foreseeable future (for example, the abfload Matlab script to read abf files). The only format that is developed in-house is the Espina seg format. We can only guarantee that this software will be readable while there are funds to maintain the software, i.e. for the duration of the HBP grant. However, Espina is freely-downloadable, and should developments in operating systems or compilers make it impossible or hard to compile from source, it would be possible to run in a Virtual machine.

3. Allocation of resources

3.1 Costs estimation for making the data FAIR:

We estimate that the cost of the extra work SP1 Partners need to make our data FAIR is 2-5 person-days on average for each dataset. This work comprises transferring data, helping SP5 with

curation, contributing to the DMP and writing better documentation, but it does not include the time SP5 partners will spend on curation or that SP7 partners will spend on infrastructure. Multiplied by the number of datasets (46), this leads to an estimate of 90-225 person days, i.e. 4.5-11.5 person-months. We will update this estimate once we have had more experience of the curation process.

3.2 Responsibilities for data management:

- SP1 partners are responsible for:
  - Producing the datasets
  - Documenting the datasets in line with SP1 and HBP standards
  - Uploading the data to HBP servers (either the Collaboratory or CSCS data buckets)
  - Obtaining necessary ethical approvals from the relevant national ethics authority
  - Working with SP5 to provide metadata for curation
- SP1 management is responsible for:
  - Maintaining the SP1 data management plan
  - Maintaining the list of SP1 datasets, checking that all links are accessible
  - Implementing HBP-wide data management policies contained In the HBP DMP in SP1
  - Where specific SP1 policies are required, deciding on SP1-wide policies or recommendations for data management, in collaboration with SP1 and SP11 partners
- SP5 is responsible for:
  - Assisting with the data curation process, including entering metadata in the Knowledge Graph
- SP11 is responsible for:
  - Drawing up the overall HBP DMP
  - Providing infrastructure to store task-specific DMP information, and generate reports from it, as requested by SP1 and other subprojects
- SP12 is responsible for:
  - Surveying SP1 partners to check whether their datasets are compliant with ethical issues as set out in the developing SGA2 Data Policy Manual written by the Data Governance Working Group

3.3 Costs and potential value of long-term preservation:

SP1 Partners do not pay directly for the costs of data preservation on HBP servers.

It is difficult to quantify the value of the data in monetary terms, though it might be reasonable to assign to the data a cost based on how much it cost to generate – in which case, the value of the data would be a sizeable fraction of the money awarded to SP1 over the course of the HBP.

The questions concerning potential uses of the data show the value in terms of utility to others inside and out with the Project.

4. Data security

When the data is on laboratory servers, individual Partners are responsible for the security of the data, subject to the regulations of their institutions. Once the data has been transferred to HBP servers (Collaboratory or CSCS data buckets) then HBP policies, as described in Section 6 of the HBP DMP, apply.

5. Ethical aspects

The SGA1 DoA covers ethical issues. Within the list of datasets, we have provided the national ethics approval authority and corresponding ID.

6. Conclusion and Outlook

The SP1 Data Management Plan has been drawn up as a key element of good data management, according to the document *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020* (Version 3.0, 26 July 2016). This DMP describes the data management life cycle for the data generated by SP1, as well as the utility of this data for modelling purposes. In addition, the DMP describes how SP1 should organize its datasets, tools and models generated internally and so facilitate further use of its outputs by other SPs.

The SP1 DMP has been adapted to the new work plan proposed for the next project phase, the SGA2, to be implemented and will continue throughout the project.

Annex 1: List of datasets, models and tools

The list of data sets (see below) will be added to this document as a pdf document generated from the excel spread sheet '200718 SP1 SGA1 Data Management Plan Information,xls'

T1.1.1: Developing the integrated FIB/SEM and SDS-FRL immunoelectron microscopy technique

T1.1.1: Nanoscale measurements of distribution of individual receptors and ion channels in cortical neurons

T1.1.2: Generation of new intrabodies / antibody fragments

T1.1.2: IACT antibody fragments for imaging

T1.1.3: Association (co-clustering) of receptors and their effector ion channels in different neuronal compartments.

T1.1.3: GluA2, GluA3 and GluN1 in CA1

T1.1.4: Electrophysiological data-hippocampus

T1.1.5: K channel kinetic and activity in a model neuron

T1.1.6: Curated list of synaptic protein-protein interactions

T1.1.6: Computational, dynamic model of LTP and LTD in a wild type Schaffer Collateral synapse

T1.1.6: KappaNEURON software package

T1.1.6: A mapping of computational models of synapses to proteins

T1.1.7: Subcellular proteomics dataset - hippocampus

T1.1.7: Synaptic Plasticity dataset - hippocampus

T1.2.1: 3D reconstructions mouse hippocampus

T1.2.1: 3D reconstructions rat hippocampus

T1.2.1: 3D reconstructions pyramidal neurons human cortex

T1.2.1: 3D reconstructions human hippocampus

T1.2.1: 3D reconstructions pyramidal neurons mouse cortex

T1.2.2: human neurons with matching Ephys

T1.2.3: The striatal microcircuit

T1.2.4: Electrophysiological data cerebellum

T1.2.5: Morphological database of major cell types of the mouse hippocampus

T1.2.5: Electrophysiological database of major cell types of the mouse hippocampus

T1.2.5: Morphological reconstructions of mouse hippocampal neurons filled in vivo

T1.2.5: Physiological characterisation of mouse hippocampal neurons recorded in vivo

T1.2.6: Database of synaptic physiological properties in the mouse hippocampus

T1.2.7: GABAergic neuron subtypes

T1.2.8: 3D digita Reconstructions of individual tlamocrtical neurons

T1.2.9: Densities and 3D distributions of synapses using FIB/SEM imaging in the human neocortex (Temporal cortex, T2)

T1.2.9: Densities and 3D distributions of synapses in the human hippocampus

T1.2.9: Densities and 3D distributions of synapses in the mouse neocortex

T1.2.9: Densities and 3D distributions of synapses in the mouse hippocampus (CA1)

T1.2.9: Immunocytochemical detection of excitatory and inhibitory terminals in the mouse somatosensory cortex by confocal microscopy

T1.2.9: Immunocytochemical detection of excitatory and inhibitory terminals in the mouse hippocampus (CA1) by confocal microscopy

T1.3.1: Whole brain interneuron distributions

T1.3.2: Wide-field imaging of cortical activity during motor learning

T1.3.2: Wide-field imaging of cortical activity one month after stroke and rehabilitation

T1.3.3: Sinthetic images for machin learining

T1.3.4: Whole brain maps of resting state brain activation

T1.3.5: 3D image of the vascular system of the mouse brain

T1.3.5: Model of intravascular and tissue partial pressure of oxygen

T1.3.5: 3D reconstruction of the vascular system of the mouse brain

T1.4.1: Analysis of micro-anatomical data

T1.4.2: Software PyramidalExplorer 1.2 for early exploratory analysis techniques for morphological data.