EBRAINS Data and Knowledge services: FAIR Data and Models for the Neuroscience Community Status at M7 (D4.1 - SGA3)

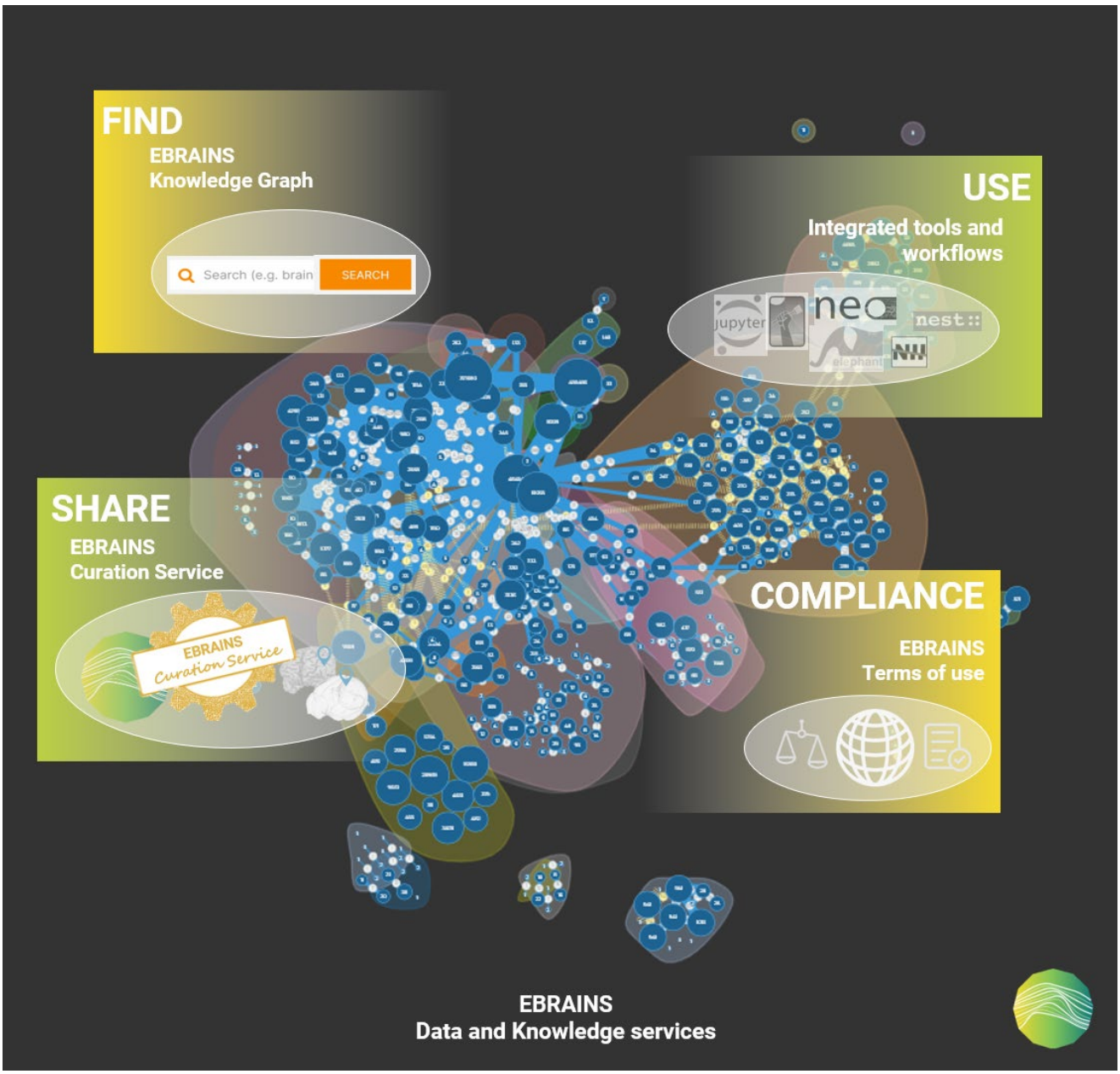

Figure 1 Key services accessible through EBRAINS Data and Knowledge

The background image depicts the graph structure of the active and continuously developing content of the EBRAINS Knowledge Graph (derived from https://kg.ebrains.eu/statistics/)

Project Number:	945539	Project Title:	HBP SGA3

Document Title:	D4.1 EBRAINS FAIR data services (SC1) status at M7
Document Filename:	D4.1 (D32) SGA3 M8 ACCEPTED 210504.docx
Deliverable Number:	SGA3 D4.1 (D32)
Deliverable Type:	Other
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA3 M7 / 30 Oct 2020
Actual Delivery Date:	SGA3 M8 / 13 Nov 2020; resubmitted 27 Apr 2021; accepted 04 May 2021
Author(s):	Jan G. BJAALIE, UIO (P81), Andrew DAVISON, CNRS (P10), Ida AASEBØ, UIO (P81), Lyuba ZEHL, JUELICH (P20), Oliver SCHMID, EPFL (P134), Mathew ABRAMS, KI (P37), Simisola AKINTOYE, DMU (P16), William KNIGHT, DMU (P16), Damian EKE, DMU (P16)
Compiled by:	Martha Elisabeth BRIGG and Roman VOLCHENKOV, UIO (P81)
Contributor(s):	Shailesh APPUKUTTAN, CNRS (P10), Onur ATES, CNRS (P10), Timo DICKSCHEID, JUELICH (P20), Tom GILLESPIE (International Neuroinformatics Coordinating Facility), Xiao GUI, JUELICH (P20), Camilla HAGEN BLIXHAVN, UIO (P81), Anna HILVERLING, JUELICH (P20), Heidi KLEVEN, UIO (P81), Stefan KÖHNEN, JUELICH (P20), Trygve LEERGAARD, UIO (P81), Elodie LEGOUÉE, CNRS (P10), Glynis MATTHEISEN, CNRS (P10), Maja PUCHADES, UIO (P81), Ingrid REITEN, UIO (P81), Ulrike SCHLEGEL, UIO (P81), Benjamin WEYERS, UT(130), Sara ZAFARNIA, JUELICH (P20), Yann ZERLAUT, CNRS (P10)
WP QC Review:	Timo DICKSCHEID, JUELICH (P20)
WP Leader / Deputy Leader Sign Off:	Jan G. BJAALIE, UIO (P81)
PCO QC Review:	Annemieke MICHELS, EPFL (P134)
Description in GA:	Knowledge Graph release with improved functionality and extended content, and inventory of related FAIR data, tools and services, prepared for, or released through the EBRAINS portal.
Abstract:	The EBRAINS Data and Knowledge services facilitate neuroscience research and discovery by providing an online conduit for both sharing and easy access to research data, computational models, and software. These services revolve around an expert-driven Knowledge Graph which combines metadata ingestion pipelines, human user input, and multiple quality assurance processes to ensure consistency and quality to aid researchers who wish to contribute. Neuroscientists looking to make their data, models, and software FAIR (Findable, Accessible, Interoperable, and Reusable) can apply for user support for curation and annotation with standardised metadata facilitating discovery and reuse by the broader research community. Moreover, clearly defined Terms of use, responsible data compliance, protection, and governance make the EBRAINS Data and Knowledge services mutually attractive for researchers both depositing and consuming content. Neuroscientists seeking data or models to support their research can find and access data and models either through a straightforward online search or a programmatic API. Moreover, these shared resources are linked to software for visualising and analysing data and for performing computerised simulations of brain function. The

	ability for scientists to more easily share, find, integrate, analyse and simulate data, accelerates the global effort to understand human brain function and disease.
Keywords:	Data sharing; data curation; data management; infrastructure; data compliance, ethics, ontology, EBRAINS.
Target Users/Readers:	computational neuroscientists, HBP/EBRAINS users, Consortium members, funders, general public, policymakers, students

Table of Contents

Table of Tables

Table of Figures

History of Changes made to this Deliverable (post Submission)

Date	Change Requested / Change Made / Other Action
13.11.2020	Deliverable submitted to EC
23.03.2021	Resubmission with specified changes requested in Review Report: There are several issues that could be improved: 1. It is not clear what has changed as the result of the migration from HBP to EBRAINS Web portals. 2. The figures reporting the increase in the size of the repository, including new datasets from SGA3 should be clarified (how many from within/ beyond HBP, how many are human data, how many are available in raw format and how many are accessible?). 3. Annex2: The Service Process model shows that there is only a one directional link between the Knowledge Graph and the Knowledge Space. Why? (It might make sense to have this bidirectional.) 4. A list of what work has been done since April 2020 is given in section 3. However also here it would be helpful to have a statement of what was available in April, what was planned, what is available now in respect to the detailed roadmap of planned features/subdeliverables listed in each task in the DoA. Analysis of identified problems, delays and risks is missing.
27.04.2021	Revised draft sent by WP to PCO. Main changes made, with indication where each change was made: 1. Result of migration from HBP to EBRAINS web portal explained (Section 2, p5) 2. Additional information added to the section reporting on the datasets (Section 3.2, p19) 3. Fixed arrow between KG and KS in Annex 2: EBRAINS Data and Knowledge services Process Model 4. Statement with reference to roadmap in the DoA and on the EBRAINS webpage added (Section 3, p18)Analysis of known problems, delays and risks added (see Annex 3: Listing of known problems, delays and risks)4.
27.04.2021	Revised version resubmitted to EC by PCO via SyGMa

1. Introduction

EBRAINS Data and Knowledge services deliver high-quality research data, computational models, and software to the research community. While several other data sharing services for neurosciences are available, none of them provide such a broad and deep set of features, covering basic and clinical neuroscience data, computational models of brain function, and software used for visualisation, analysis, and simulation.

To achieve the greatest impact, the EBRAINS Data and Knowledge services have embraced a set of data sharing principles. These principles emphasise that datasets should be Findable, Accessible, Interoperable and Reusable (FAIR) to ensure maximum use and reuse of the shared data (Wilkinson *et al*, Sci Data 2016, https://doi.org/10.1038/sdata.2016.18).

For scientists wishing to share data, the services include data curation and data storage, DOIs for citation of the data, defined conditions and licenses for use of the data. For scientists wishing to find and access data, both an intuitive user interface and programmatic API are available, with both free-text and structured search based on extensive metadata annotations. Datasets, models and software descriptions have links for direct download, inline preview and links to viewers for exploration of the data, and links to EBRAINS workflows for data analysis.

A data, model, and software curation process delivers metadata enrichment and quality assurance. Depositors (contributors) are provided with intuitive tools and services for managing their metadata and, when relevant, for extending the metadata repertoire. The metadata management support builds on metadata standards and ontologies that are developed and maintained by EBRAINS in collaboration with other brain initiatives and the international neuroscience community. It thus ensures the highest possible level of FAIRness. The curation process also includes routines to ensure that all data shared via EBRAINS comply with applicable EU and national regulations, as well as EU-supported ethical principles. Persistent data storage is delivered for open data as well as for data requiring protection. Requests for access to data in the protected storage will be facilitated by application procedures available through the data landing pages (data cards in the EBRAINS Knowledge Graph), covering also the setting up of the necessary inter-institutional agreements for use of the data.

The number of FAIR datasets, computational models, and software tools made available in the EBRAINS Knowledge Graph is continuously growing. With the continuing and expanding effort through EBRAINS, the services provided will be increasingly relevant and attractive for users in the neuroscience community in academia as well as industry.

2. Current status - What is available now

The EBRAINS Data and Knowledge services have been available to the community through the Human Brain Project web site since April 2018, in the beginning embedded in the HBP Project website. In October 2019, they were migrated to the EBRAINS web portal and presented as an integrated suite of services, policies, and practices for FAIR data. The changes introduced included 1) a new Share data service, targeting the broader neuroscience community, and with extensive support also users outside of the HBP, 2) the inclusion of models in addition to research data, and 3) an improved user interface and usability for the key product: the EBRAINS Knowledge Graph. The new EBRAINS Data and Knowledge services is thus targeting the broader neuroscience community, making it easy for users to share and publish data and models, and to find heterogeneous data, models, and related software relevant for a broad range of research fields in basic and clinical neuroscience.

The services included are:

- EBRAINS Knowledge Graph
- EBRAINS Curation
- EBRAINS Compliance Management, Data Protection, and Data Governance
- KnowledgeSpace

Among these services, the first three are tightly integrated, whereas KnowledgeSpace is a separate service, supported by EBRAINS with the aim of connecting the EBRAINS Knowledge Graph to a broad range of other services worldwide. Below, we provide a summary of the features that are currently available in each of the services (October 2020) with a focus on what they provide to the end users.

2.1 EBRAINS Knowledge Graph

The EBRAINS Knowledge Graph ((KG), https://kg.ebrains.eu) is a metadata management system that provides fundamental services for making neuroscientific data, models, and related software FAIR.

The knowledge graph technology is used in many organisations and for many purposes. It organises and links information from various sources in a web-like structure for the purpose of defining and exploring relationships and generating knowledge. Knowledge graphs have properties that make them more suitable than conventional fixed-schema databases in many application areas:

- they allow existing data models to be extended at any point in time with new properties and connections
- they can be used to generate a graphical representation of the relationships between any of its data points, allowing dynamic and flexible navigation of the content
- complex relationships between data points can be represented more easily and thanks to semantic annotations, new connections can be inferred

Adopting the knowledge graph technology to a huge and diverse field of research such as neuroscience does not come for free. In many other application areas, pre-processed information is fed into the graph structures via automated ingestion pipelines. For the field of neuroscience, such pipelines are in most cases not easy to establish since the pre-processed information is either not available or not possible to interpret for others than those who generated the information. The EBRAINS Knowledge Graph addresses this challenge through an “expert-driven” approach. Ingested data are controlled by manual, semi-automated and automated quality assurance processes allowing domain-specific experts (in this case the neuroscientists of the EBRAINS Curation services) to actively validate and ensure conventions and consistency of the data. Neuroscience-specific metadata standards and conventions are introduced while supporting extensions and adaption for future needs. The design employed makes conflicting information transparent and therefore enforces scientific discussions. This approach (which we call “expert-driven”) adds several challenges which are addressed by various components built on top of the core services of the Knowledge Graph.

2.1.1 Key components

See Table 1: Key components of EBRAINS Knowledge Graph

Table 1: Key components of EBRAINS Knowledge Graph

Key Components	Description
Core	KG core is the central component of the EBRAINS Knowledge Graph. It provides the required APIs to ingest and consume the data into and from the graph. It provides the linking of instances, provides fine-grained access-control, and has inbuilt reconciliation and inference logic.
Editor (Figure 2)	The KG Editor is a web application allowing experts to interact (read / edit) directly with the underlying graph structures via a convenient user interface. The interface reduces the underlying complexity of the graph structure by contextualising it to the curation workflows. It is closely integrated with the quality assurance and review process, e.g. by allowing to invite other users for review and to persist reviewed and quality assured states of an instance by staging them to a “released” snapshot.

Query Builder	A query API with a descriptive query language queries allows the programmatic clients to define exactly which subset of the data structures they are interested in. Queries can be saved to build client-specific, custom API endpoints. Additionally, queries are used by the system internally to reflect on the strength of dependencies between instances and helps to optimise other internal tools (e.g. the KG Editor). The Query Builder provides a visual tool allowing to create and combine queries by picking the available fields in the underlying graph structure.
Search UI (Figure 3)	The most visible component of the EBRAINS Knowledge Graph is its Search User Interface. It aggregates information from the graph structure into visual cards (Data cards, Figure 4) representing the data in a convenient way for end-users. The UI provides full-text search capabilities, faceted search mechanisms and allows to navigate between dependent resources. It is tightly bound to the DOI registration mechanism used to make data citable. The UI also allows to navigate to additional EBRAINS services such as 3D viewers to facilitate exploring of the data.
Statistics	KG Statistics provides a visual overview of the overall graph. It is used to explore the underlying structures and to detect potential irregularities.
Automation	A set of automation scripts handle the automated ingestion of data from internal and external sources (e.g., information about file repositories and ontologies), integration with external services (e.g. DOI registration and synchronisation), automated validity checks and similar.
Metadata monitoring	The monitoring component feeds information about the metadata in the graph into a GitLab issue tracking system to report on the current state and potential issues. It builds an interface between the overall curation workflow and the EBRAINS Knowledge Graph and simplifies the organisational aspects of the metadata management.

2.1.2 Associated tools

EBRAINS provides tools for visualising, analysing and otherwise reusing datasets and their metadata. In many cases these are provided through other EBRAINS Service Categories, such as the viewers for image data and atlases provided by the EBRAINS Atlases services. Others are provided directly by EBRAINS Data and Knowledge services. For example, Neo Viewer (Figure 5) provides web-based visualisation of electrophysiology data, with support for a wide range of neurophysiology file formats. The KG Query API enables programmatic querying of the Knowledge Graph content and access to datasets. Python libraries to facilitate working with the Knowledge Graph APIs are available for download and in the Jupyter notebooks provided by the EBRAINS Lab service (Figure 6).

2.1.3 User benefits

The benefits offered to the individual users of the EBRAINS Knowledge Graph include:

- Access to quality-assured metadata describing datasets, models and software in human- and machine-readable ways
- Search User Interface for full-text queries and faceted filters on top of the metadata, allowing the user to narrow down the search results by text patterns and regular expressions as well as spatial location in the brain
- Query building and storage of queries, allowing the user to select types, fields and relations of interest and execute the queries over again through an API or automatically generated python code
- Direct download of the registered data
- Precise description of the conditions of use of the data, models, and software, including licenses and information on how to cite

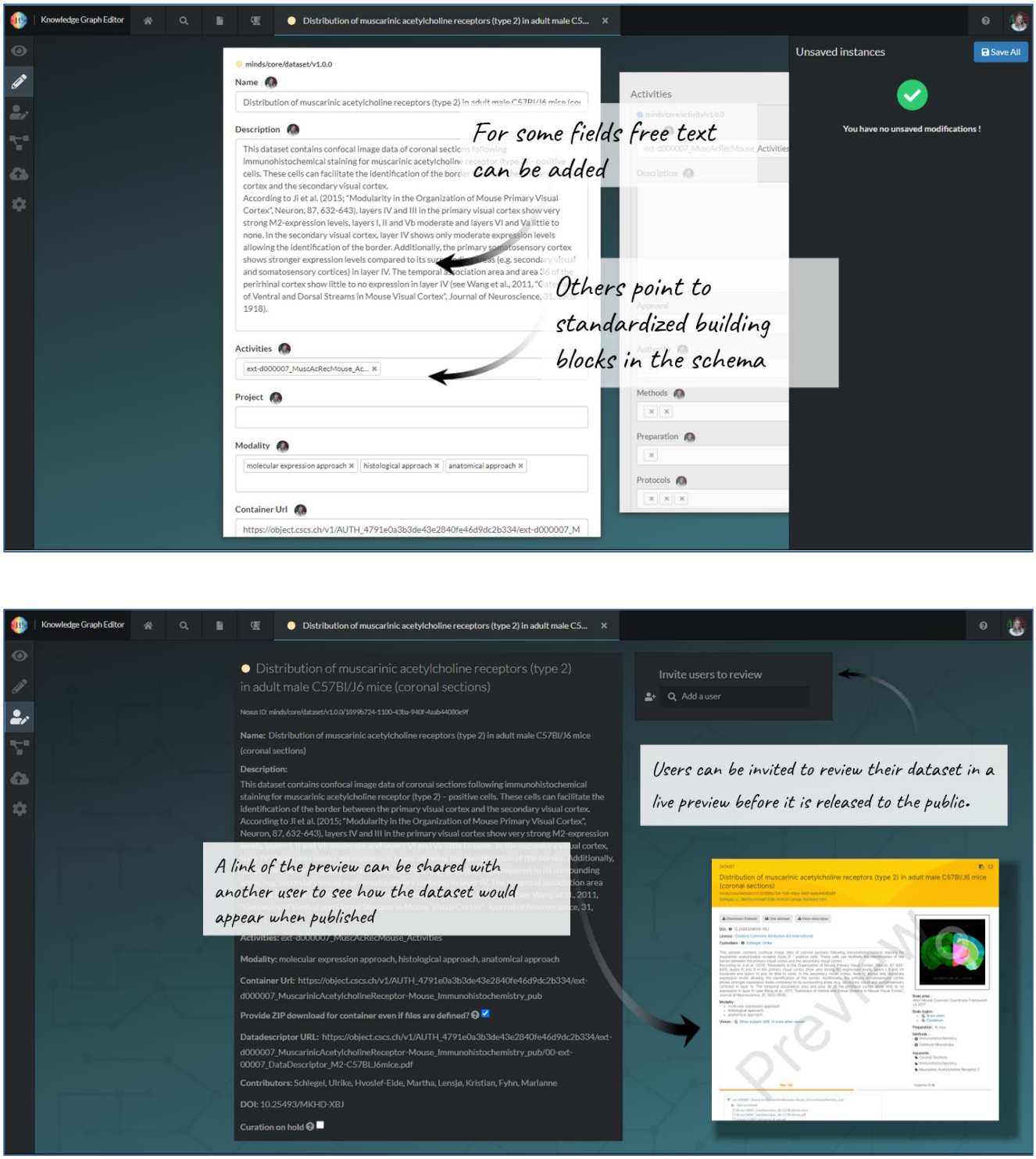

Figure 2: The EBRAINS Knowledge Graph Editor

The EBRAINS Knowledge Graph Editor is a tool developed to support the curation process and publication of datasets, models, and software to the EBRAINS Knowledge Graph. It makes the curation process more automated and less time consuming. The current version of the Editor is used by the curation team whereas future versions will also be available for the providers of data, models, and software.

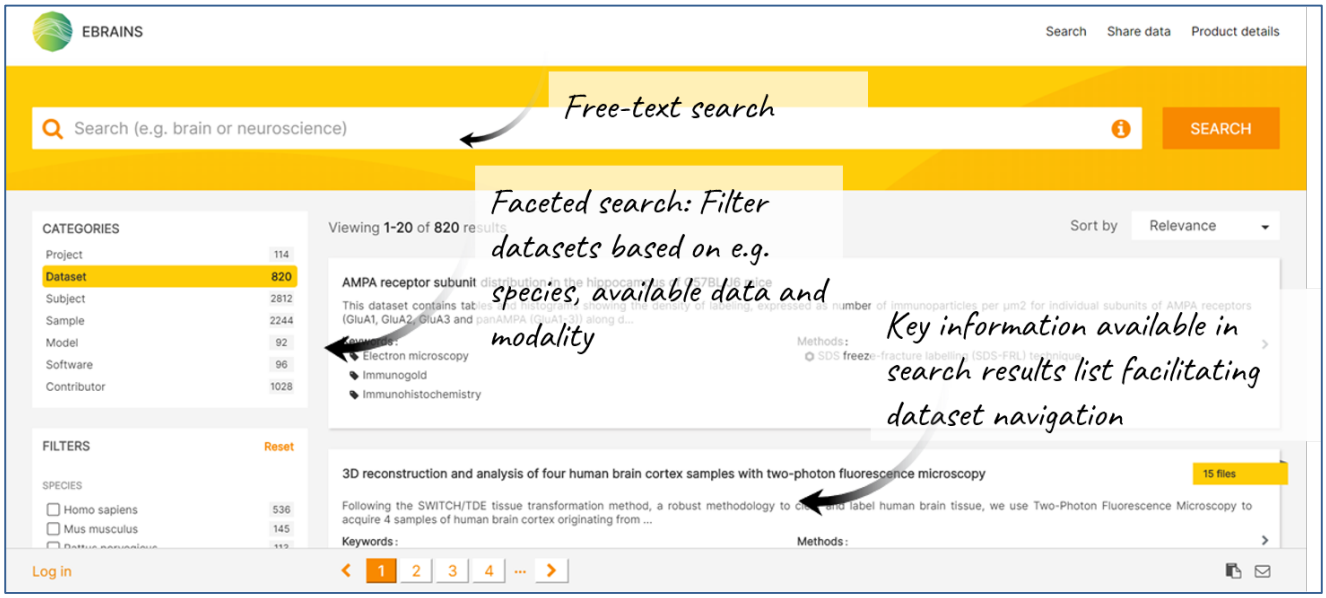

Figure 3: The Knowledge Graph Search User Interface

The Knowledge Graph Search User Interface (https://kg.ebrains.eu/search/) has been continually modified to create an easier search for the researcher seeking to find data, models, or software. It allows for free text searches and provides filters to narrow searches based on metadata such as type of method, type of data, and species.

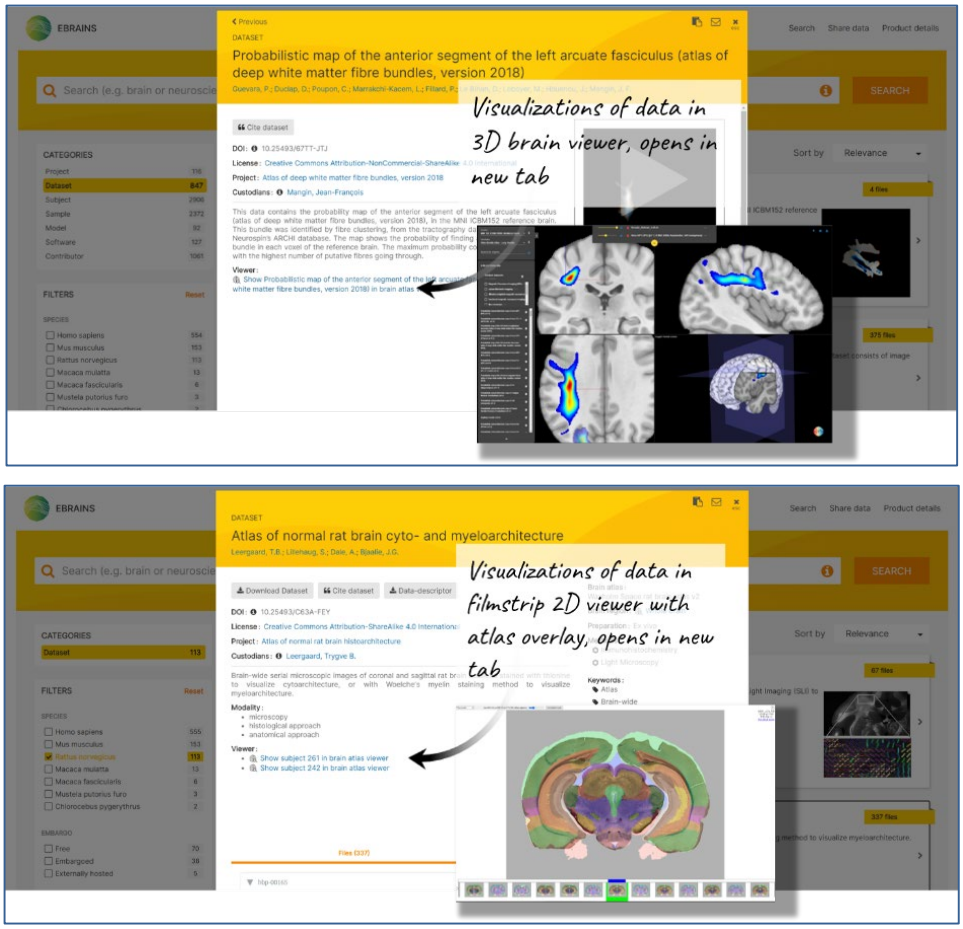

Figure 4: Dataset cards in the EBRAINS Knowledge Graph.

The Dataset cards contain key information about the dataset, e.g. its metadata, terms of use, and how to cite and reuse the data. Datasets that have undergone spatial registration (anchoring) to an atlas contain links to hosted viewers for 3D representations (coordinate-based location) and 2D (actual images) of the data.

Links to the Dataset cards shown:

https://kg.ebrains.eu/search/instances/Dataset/7e227fef-dc93-4593-a0ea-707996992dd8

https://kg.ebrains.eu/search/instances/Dataset/6ce1f96ae210b2335b75a793367e3865

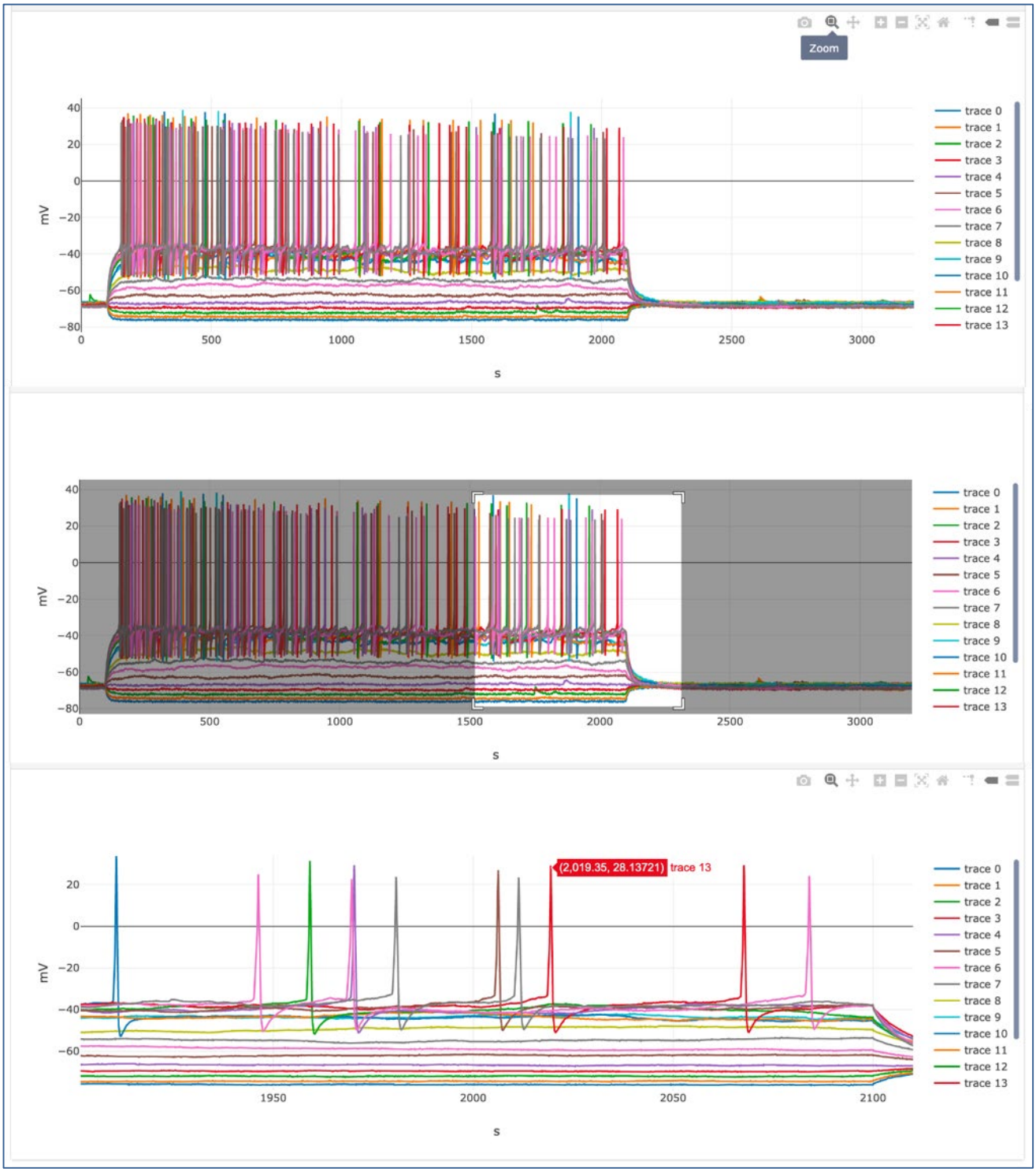

Figure 5: Neo Viewer

Neo Viewer provides web-based visualisation of electrophysiology data, with support for a wide range of neurophysiology file formats. It is available as an EBRAINS web service and can also be self-hosted. Electrophysiology traces can be zoomed, scrolled, and saved as images. Individual points can be measured off the graphs.

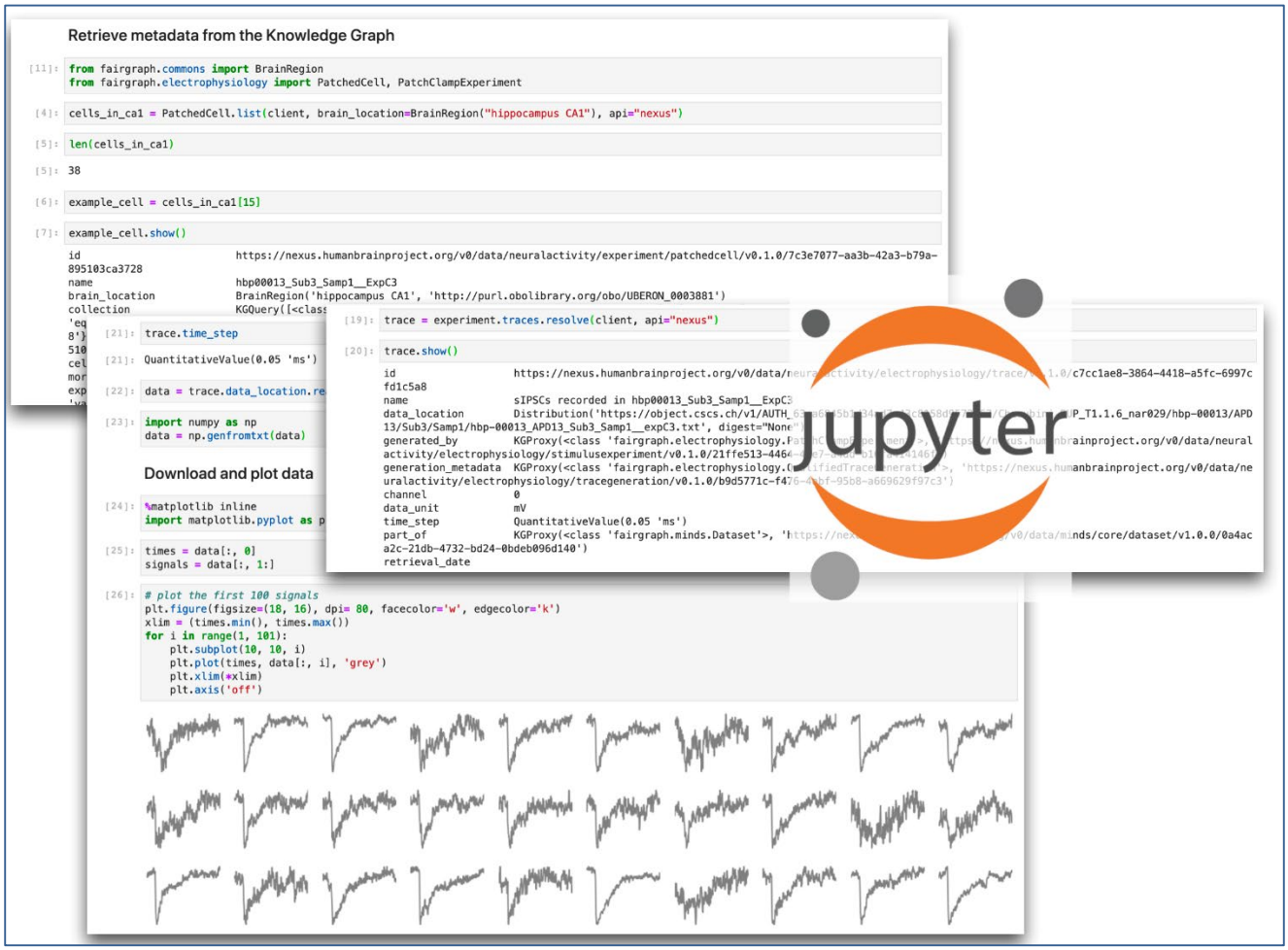

Figure 6: The Knowledge Graph API in Jupyter notebooks

The Knowledge Graph APIs can be used to download, analyse and visualise both data and metadata in Jupyter notebooks, like those provided by the EBRAINS Lab service.

2.2 EBRAINS Curation

EBRAINS Curation oversees the deposition of data, models, and software in the EBRAINS Knowledge Graph, provides assistance to the depositors, and develops and maintains the metadata schemas required for making data, models, and software FAIR.

While the term data curation may have different connotations and meanings, the term generally refers to processes described as adding value to data. A broad definition typically emphasises the management of data through the lifecycle and the integration of data into repositories, enabling discovery, retrieval, reuse and maintaining quality. Data curation is often an activity in the domain of librarians and information scientists. With the emergence of repositories for scientific research data, there is a need for more expert-driven curation, addressing domain-specific requirements, and including the integration of computational models and software tools.

EBRAINS Curation delivers curation of data, models, and software relevant for the broader domain of neuroscience. The curation team works closely with researchers and developers across many disciplines. The responsibilities of EBRAINS Curation include:

- Developing and maintaining the curation workflow from submission through a series of steps resulting in FAIR research data, models, or software available through the EBRAINS Knowledge Graph
- Developing, maintaining and extending metadata schemas and ontologies, relevant for neuroscience data, models, and software

- Support and guidance of neuroscientists and developers of neuroscience data, models, and software through the curation workflow, leading to the assignation of metadata and the sharing of their work
- Ingesting metadata in the Knowledge Graph; making data, models, and software available with DOIs and licenses for use
- Monitoring the incoming requests and existing data, models and software entries in the Knowledge Graph, evaluate workload/relevance of novel entries and ensure/optimise the quality of the database content.
- Engaging with the community regarding EBRAINS Curation Services to inform about Open data, curation procedures as well as metadata schemas, standards, and general good practices for metadata and data management.

An overview of EBRAINS Curation in relation to the Knowledge Graph and associated tools is shown in Figure 7.

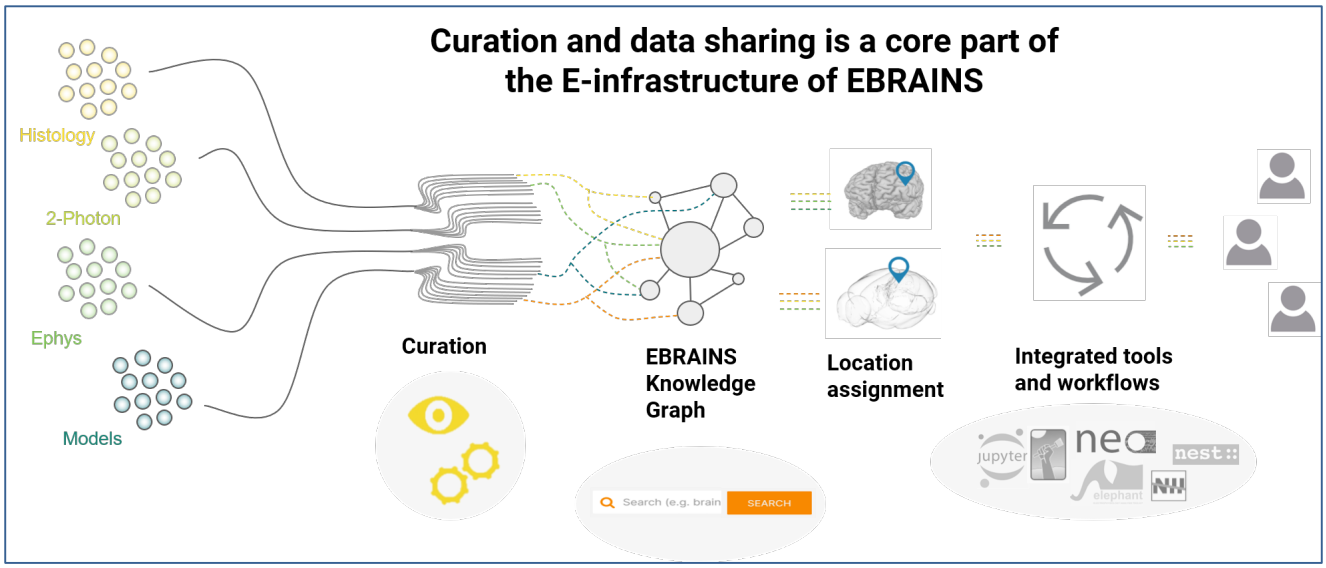

Figure 7: EBRAINS Curation, Knowledge Graph, and related tools

EBRAINS Curation provides the basis for organised and managed content in the EBRAINS Knowledge Graph, connected to tools and workflows for analysis of Knowledge Graph content.

In order to deliver on these responsibilities, EBRAINS Curation has the following components:

2.2.1 Key components

Metadata schemas

Currently, the metadata schemas employed in the curation process are called the HBP Minimum Information for Neuroscience Data Sets (HBP-MINDS). These schemas capture essential information, such as study-target, subject category, tissue-samples, techniques used, related publications, and person contributing the data. From November 2020, HBP-MINDS (v1.0/2.0) will be replaced by openMINDS (v3.0).

Open Metadata Initiative for Neuroscience Data Structures (openMINDS)

A large investment has been made in developing the new and much more comprehensive set of metadata schemas: the open Metadata Initiative for Neuroscience Data Structures (openMINDS). openMINDS is an open source, community-driven project hosted on GitHub (https://github.com/HumanBrainProject/openMINDS) comprising a set of metadata schemas collections for increasing the findability, accessibility and reusability of data repositories down to single files that originate from various neuroscience modalities and species. The openMINDS GitHub is managed and maintained by a core development team with members from EBRAINS Curation, Knowledge Graph and KnowledgeSpace teams. The next version of metadata schemas,

openMINDS core (version 3.0 https://github.com/HumanBrainProject/openMINDS_core) will be released in November 2020, with integration into the Knowledge Graph starting also in November 2020.

All openMINDS metadata schemas are designed to be used as architectural building blocks for graph databases, such as the EBRAINS Knowledge Graph. The modular structure of the schemas allows users to easily establish cross links between registered research products (datasets, models, software) within such databases. Furthermore, the openMINDS schemas allow users to define relations to ontologies, providing the possibility to connect data beyond the corresponding database to resources hosted elsewhere.

Extensions to openMINDS core

The Spatial Anchoring of Neuroscience Data Structures (SANDS) schema is an extension created to standardise the identification of the anatomical location of neuroscience data, a major step towards close integration of KG with EBRAINS atlas services. SANDS is hosted on GitHub (https://github.com/HumanBrainProject/openMINDS_SANDS) and will be released in November 2020 (version 1.0). It employs the EBRAINS brain atlases and the formats and requirements for the tools used to visualise and analyse the data.

In addition to SANDS, other in-depth schema collections will extend the openMINDS core and allow information to be captured that is useful for automated analysis pipelines or simply in defining the full provenance of the research products (see Curation workflow, below).

Ontologies

EBRAINS employs internationally accepted ontologies to facilitate metadata standardisation in an effort to align with global neuroscientific nomenclature practices and to make the intricate and established relationships within and between neuroscientific data and metadata accessible.

The enrichment of existing ontologies is performed as a collaborative workflow between the EBRAINS curation team and the US Neuroscience Information Framework (NIF). This collaboration facilitates alignment of terminologies with other projects (e.g. the US BRAIN initiative, and INCF Neuroshapes). The development of ontologies is an ongoing and continuous process. The EBRAINS Knowledge Graph pulls the latest relevant information from the NIF ontology services on a daily basis, including:

- Neuron Phenotype Ontology development
- Atlas ontology alignment for openMINDS and SANDS
- Methods ontology (https://github.com/SciCrunch/NIF-Ontology/tree/methods)
- Parcellation ontology (https://github.com/SciCrunch/NIF-Ontology/tree/parcellation)
- Neuron ontology (https://github.com/SciCrunch/NIF-Ontology/tree/neurons)

Curation workflow

The EBRAINS curation workflow is a well-documented and consistent process for curation and metadata management (https://ebrains.eu/services/data-knowledge/share-data/). It is delivered as a service to all contributors of data, models, and software. The workflow begins with the request for curation submitted by the contributor and ends with FAIR metadata and research data, models, and software available through the EBRAINS Knowledge Graph. It is divided into two distinct activities: support to the contributors and a systematic review before data and metadata are made discoverable in the Knowledge Graph. The contributors complete a structured web form to describe the dataset, model or software. The curation team then evaluates, quality controls, revises and enters structured metadata into the Knowledge Graph, checks that the experiments had appropriate ethical approval, and assists with data upload to archival storage. The resulting published datasets adhere to FAIR principles and meet the requirements of funders and of publishers such as Springer Nature.

The multiple steps are explained on the EBRAINS Share web pages (https://ebrains.eu/services/data-knowledge/share-data/) and in the EBRAINS Curation Collab (Figure 8, https://wiki.ebrains.eu/bin/view/Collabs/data-curation).

As outlined below, all datasets undergo a basic curation. Through the Atlas integration, the majority of datasets are also annotated with spatial metadata, situating the data source within the brain atlas for the relevant species. Certain datasets also undergo further, in-depth curation, adding more fine-grained metadata where needed for evaluation and reuse of the data.

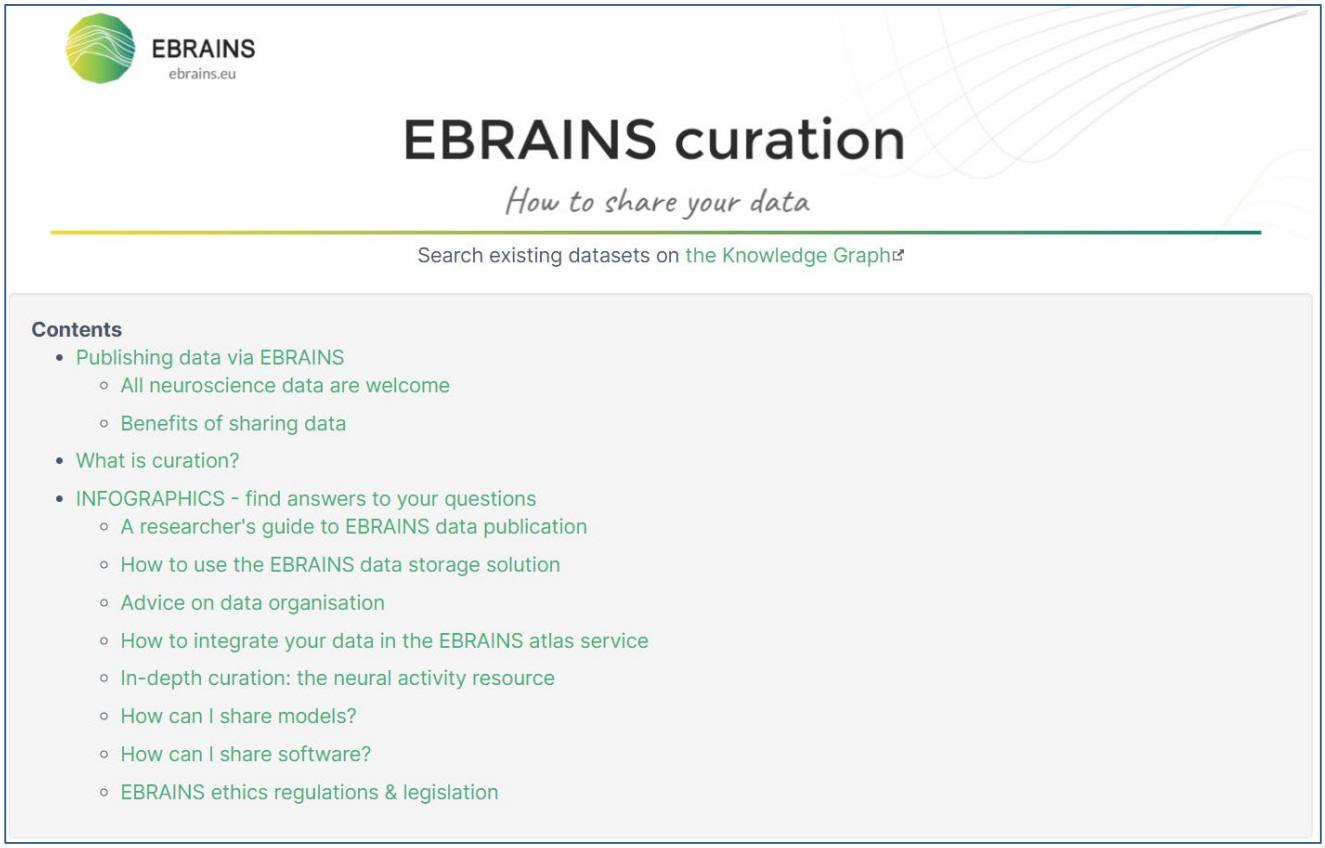

Figure 8: EBRAINS Curation Collab

The publicly accessible EBRAINS Curation Collab, containing detailed information about the curation process and relevant background information, targeting depositors and consumers of content in the EBRAINS Knowledge Graph.

Basic curation

The Basic curation step currently refers to request evaluation, the MINDS schema, a Data Descriptor, interactive dialogue with the depositor and the data provision itself. All content (currently 854 datasets, 92 models, 128 software) in the EBRAINS Knowledge Graph has been subject to basic curation resulting in metadata listed in the faceted search (left column in the EBRAINS Knowledge Graph Search UI) with number of datasets matching specific combinations of metadata.

From November 2020, this curation step will be using the new openMINDS metadata schemas (see above), thereby providing increased findability and interpretability through inclusion of richer information.

Atlas integration

The Atlas integration step introduces semantic or spatial metadata serving to identify the anatomical location of neuroscience data, at the level of names of brain structures (using the atlas ontologies for structure names) or spatial coordinates (using the atlas coordinate frameworks). The enhanced metadata added by this curation step improves findability and reusability: for example, the metadata can enable identification of the datasets available for a brain structure or the brain-wide distribution of selected features in a particular dataset. Whereas the semantic metadata are findable through the Knowledge Graph Search UI, the spatial metadata are available in the data cards. The technology required for spatial search has been developed and is considered for implementation as a functionality in the EBRAINS Atlases services.

As of October 2020, atlas integration for human data is available for a selection of key datasets which are presented through the EBRAINS Atlases services. For rodents, 261 rodent datasets have been semantically integrated in the rat/mouse atlases and can be identified through a search in the Knowledge Graph: https://kg.ebrains.eu/search/instances/Dataset/7246bb9d-5f8c-4281-a142-ab931f824f9c. Of these, 109 datasets have also been spatially integrated in the rat/mouse atlases.

From November 2020, this curation step will use the Spatial Anchoring of Neuroscience Data Structures (SANDS) schema (see above).

In-depth curation

In-depth curation adds information about electrode configurations and properties, information about the experimental preparation (culture, slice, or *in vivo*), and detailed experimental protocols including visual, behavioural, or electrophysiological stimuli. The enhanced metadata added by this process improve findability and reusability: for example, the metadata can enable automated reconstruction of stimulation and recording protocols for validation of computational models with respect to the experimental data. The in-depth metadata are not yet visible in the Knowledge Graph Search UI, but can be accessed programmatically through the fairgraph Python API. Modality-specific features described by the in-depth curation schemas can be used to search for and filter datasets. This grants data consumers direct access to specific subsets of information, thereby increasing the efficiency of searches and providing data tailored for study, citation, reuse, and analysis. A new version of the Neural Activity Resource app, giving graphical search and visualisation of the in-depth metadata will be made available in the EBRAINS Collaboratory during the second half of 2021.

As of 1st October 2020, 44 datasets included in the EBRAINS Knowledge Graph have undergone in-depth curation, broken down by recording modality as follows:

- 32 patch clamp
- 3 optical imaging
- 3 extracellular electrode
- 3 morphology reconstructions
- 2 EEG
- 2 ECoG

From November 2020, the developed in-depth metadata schemas will be integrated into openMINDS as formal extensions (see above).

2.2.2 User benefits

EBRAINS Curation is a key part of the EBRAINS Data and Knowledge Services. It contributes to making EBRAINS Data and Knowledge Services different from the generalist repositories for data sharing (for example Zenodo, Figshare, OSF) and also different from many discipline-specific repositories that are often not curated in the same way. The journal Nature Scientific Data has included EBRAINS as a recommended repository, in line with its guideline: “data should be submitted to discipline-specific, community-recognized repositories where possible".

Overall, the benefits offered to the individual users through curation as an integral part of the EBRAINS Knowledge Graph include:

- Active curation, quality control
- Assistance to the data providers from the curation team
- Much richer, neuroscience-specific metadata
- Increased FAIRness of data, models, and software, including more powerful search capabilities and increased reusability (enabled by the rich metadata)

- Integration with other EBRAINS services, such as highly parallel data analysis pipelines, modelling and simulation tools

Curation can be requested through the curation request form available on the EBRAINS Share web page (see Curation workflow, above).

2.3 EBRAINS Compliance Management, Data Protection, and Data Governance

EBRAINS Data and Knowledge Services are underpinned by legal and ethical principles, which are built into the design of the services through compliance management, data protection, and data governance.

EBRAINS Ethics Compliance, Data Protection and Data Governance activities collect a number of processes, policies and responsibilities which cut across the various services on EBRAINS platforms. Broadly, they ensure that activity on EBRAINS complies with legal and ethical standards. Ethics compliance refers to adherence to ethical laws and standards. Data Governance is the overall management of the availability, usability, integrity and security of data. It encompasses the principles and practices that signify acceptable processes of the collection, storage, processing, curation, use, and deletion of data. Data Protection refers to the legal process of safeguarding information relating to identified or identifiable individuals from breach, compromise or loss. Compliance management, Data Protection and Data Governance activities are conducted by an interdisciplinary team of ethicists, lawyers and data governance experts, but they involve collaboration across the infrastructure with technologists, engineers and scientists. EBRAINS Compliance Management, Data Protection, and Data Governance services are monitored, updated and improved on a regular basis. The aim is to demonstrate good practice by focusing on balancing the competing interests of research and innovation with privacy and data protection.

The EBRAINS Compliance Management, Data Protection, and Data Governance services include:

- Ethics Compliance: Working with researchers and data providers to ensure ethical compliance of all human and animal data integrated into the EBRAINS infrastructure, and that evidence of acceptability is available to appropriate parties using the data. Ethics compliance checks are a key part of the data curation process, each human and non-human animal dataset is reviewed for compliance with EU law and Horizon 2020 ethical standards[1]. All data shared through on EBRAINS platforms should, therefore, be compliant with EU law.
- Data Protection: Monitoring compliance to the General Data Protection Regulation (GDPR)[2] and any other applicable data protection laws, supporting researchers, providing education, and drafting data protection compliance routines throughout the project. The Data Protection Officer (DPO) assists with monitoring of internal compliance and data protection obligations across the project including accountability to data subjects, consultation on data processing activities and providing advice and recommendations on compliance with applicable laws.
- Data Governance: Coordinating and leading the data governance activities and research, with a particular emphasis on ethics-related aspects. Insights derived from the latest data governance and data ethics research informs EBRAINS activity. These insights are disseminated throughout the infrastructure via the EBRAINS Data Governance Working Group: a cross-cutting working group from all areas of the Project. EBRAINS Data governance activity has already produced a number of key policies and documents through the Data Governance Working Group (DGWG), including:
    - The Data Use Agreement[3]: this document governs the use of pseudonymised human data made available through EBRAINS.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm
[2] https://gdpr-info.eu/
[3] https://strapi-prod.sos-ch-dk-2.exo.io/EBRAINS_Data_Use_Agreement_90858e7836_ef3ee29d50.pdf

- The EBRAINS access policy[4]: This access policy is aimed at facilitating the responsible, transparent, extensive and appropriate sharing, access and use of the services and resources provided through EBRAINS.
- EBRAINS General terms of use[5] that define the relationship between EBRAINS and individuals that access and use EBRAINS Resources (data, models, software), tools and services.

- Data Management: Implementing insights derived from this research to develop ethics and data protection compliance processes for implementation in the EBRAINS infrastructure. Data management involves having a comprehensive overview of the data found and shared on EBRAINS platforms, this allows the targeting of key areas of concern where additional support can be provided.

2.3.1 User Benefits

EBRAINS Compliance Management, Data Protection and Data Governance ensure that any data shared to users through EBRAINS platforms is ethically and legally acceptable. EBRAINS users can be confident that:

- All data shared through EBRAINS has been subject to an ethics compliance check. EBRAINS aims to provide access to legal, ethical and responsible data. EBRAINS activities are informed by cutting-edge data governance and data protection research. This ensures that EBRAINS services are provided to users based upon the latest insights into data ethics, data governance and data protection laws and standards.
- EBRAINS policies and practices ensure that any data subject access requests are adequately addressed. Relevant questions or queries related to Data Protection can be addressed to the EBRAINS Data Protection Officer.

Comprehensive details on EBRAINS terms and policies can be found on the EBRAINS website[5].

2.4 KnowledgeSpace

KnowledgeSpace (http://knowledge-space.org), is a community-based, data-driven encyclopaedia for neuroscience that provides a unique, global interface between current brain research concepts and the data, models, and literature that support or weaken their definition.

KnowledgeSpace is a knowledgebase framework that functions to extract, accumulate, organise, annotate, and link information to datasets and models that comply with the FAIR principles. The KnowledgeSpace approach is to ontologically link descriptions/definitions of neuroscience concepts to related data, models, and literature entries found in EBRAINS Knowledge Graph and 13 other leading neuroscience repositories, as listed below.

- Cell Image Library
- Allen Institute
- Open Source Brain
- NeuronDB
- ModelDM
- NeuroLex
- PubMed
- Blue Brain Project

[4] https://strapi-prod.sos-ch-dk-2.exo.io/EBRAINS_access_policy_v1_1_ed4b84ee9e_e90146495f.pdf
[5] https://ebrains.eu/uploads/EBRAINS_General_Terms_of_use_e457353c1a.pdf

- Neuromorpho
- Brain Biodiversity Bank
- GENSAT
- NeuroML
- NeuroElectro

Ingestion of metadata into KnowledgeSpace links data and models to brain research concepts and PubMed entries related to the concept, and integrates them into a larger ecosystem of data and models, while maintaining access control and repository curation standards. Being indexed in KnowledgeSpace is a proof of adherence to the FAIR principles, since only repositories adhering to the FAIR principles are discoverable through KnowledgeSpace. Metadata for EBRAINS Knowledge Graph data and models have been ingested into KnowledgeSpace, thus making Knowledge Graph data and models discoverable through KnowledgeSpace search. It should be noted that, while Knowledge Graph data and models are discoverable via KnowledgeSpace search, the data and models in the Knowledge Graph have not yet been fully integrated into the KnowledgeSpace ecosystem. Full integration of Knowledge Graph data and models into KnowledgeSpace will follow with the implementation of openMinds core and SANDS from November 2020 (see above).

2.4.1 User benefits

- KnowledgeSpace provides users with Wikipedia-like descriptions of neuroscience concepts that are (ontologically) linked to data/models from 14 of the world's leading neuroscience data repositories, including EBRAINS, and to related PubMed entries
- KnowledgeSpace provides a search interface where users look up their term of interest and receive a description of that term linked to anatomy, expression, morphology, and physiology data and models (over 1.6 million files) available for reuse to support computation models, analysis, or use as sample datasets
- KnowledgeSpace provides a single entry where users can find open, publicly available data and models

3. Recent Developments

The EBRAINS Data and Knowledge services have been extensively developed since April 2021, following the roadmap for the beginning of the current phase of the HBP. Further roadmaps for these services are shown on the service pages: https://ebrains.eu/services/data-and-knowledge#services.

3.1 EBRAINS Knowledge Graph

Developments since April 2020 have focused on improvements in the Knowledge Graph Search:

- In order to address sustainability, the existing service has been updated to the newest versions of its underlying technologies and prepared for proper horizontal scaling, which allows to respond to increased use and to therefore quickly react to performance and/or availability issues. The users will therefore profit from a more reliable service.
- Optimisations in the UI (hierarchical file view / download sub directories as ZIPs), as well as new facet filter mechanisms, e.g. for methods and keywords, as well as the introduction of new aspects (modalities), improve the usability of the system in both, finding data and accessing them. Additionally, the hierarchical file view is a required first step for additional features such as cross-linking files with compatible software and similar.
- Access to protected areas ("in progress") have been optimised to simplify reviewing processes and visual markers have been introduced for improved clarity on the shown state.

In addition, improved data analytics on top of the Knowledge Graph has helped to detect possible irregularities in the curation process and therefore to reduce the number of errors in the data. This information is directly fed into the curation process. The users profit from a higher quality of data.

3.2 EBRAINS Curation

Since April 2020, there has been a significant increase in the number of datasets, models, and software curated and included in the EBRAINS Knowledge Graph, as shown in Figure 9. The amount of data originating from data providers external to the HBP increased gradually during 2020 to a level of 10 % in October 2020. Data originating from human subjects made up 65 % of the datasets, rodent data 30 %, and data from other species 5 %. Data providers often request that data are kept under embargo until publication of a related journal article. More than 80 % of the data were nevertheless openly available in October 2020. The large majority of datasets contained primary (and not processed) data. The models were primarily of HBP origin, while for software included in the Knowledge Graph, around 50 % originated from outside the HBP.

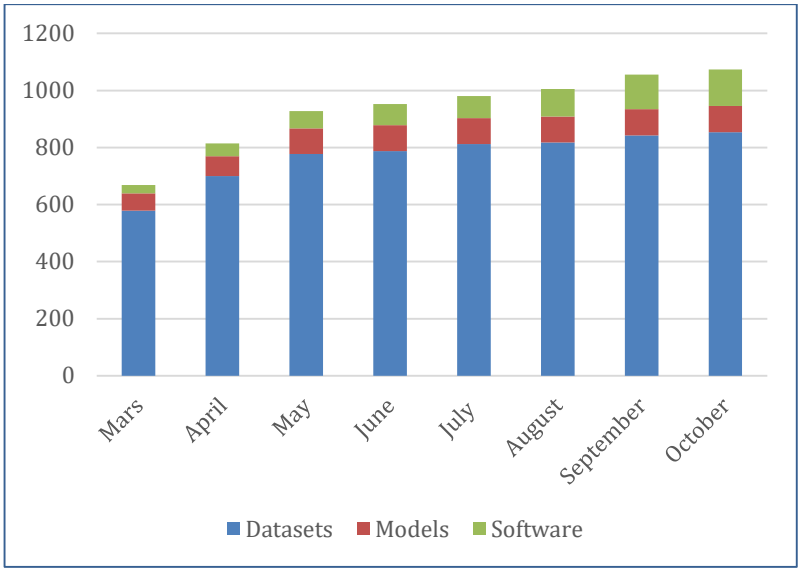

Figure 9: EBRAINS Knowledge Graph increased use

Recent increase in number of datasets, models and software integrated in the EBRAINS Knowledge Graph (March-October 2020)

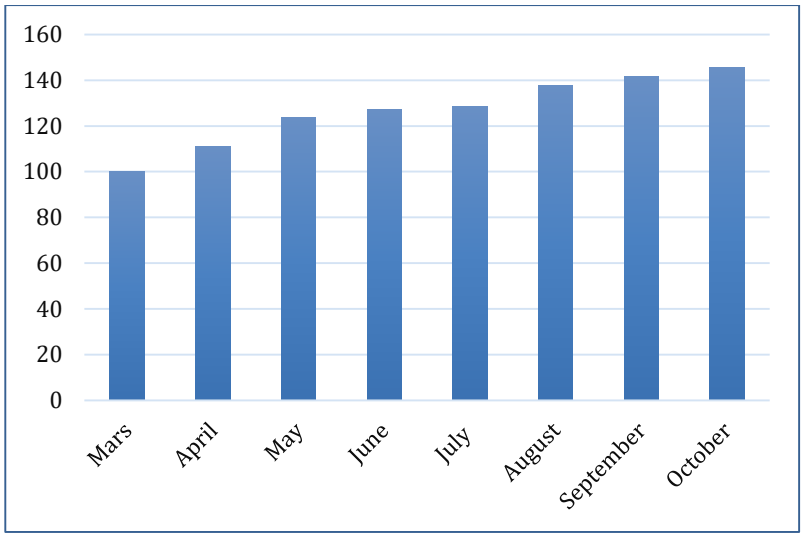

Figure 10: Index of curated data (ICD)

The index is composed of number of *subjects* and number of *samples*, each given the same weight. These parameters express the increasing amount of data better than datasets (since datasets range from small, with few subjects and/or samples, to large, with many subjects and/or samples). ICD value 100 corresponds to the number of subjects (1,740) and number of samples (1,960) recorded at the end of March 2020.

The increasing amount of data measured through an index based on samples and subjects is shown in Figure 10

Developments since April 2020 include:

- A new automated internal error tracking system has been implemented: Due to the growth in the number of datasets in the Knowledge Graph and the development of novel features of the curation workflow and Knowledge Graph, the curators are to an increasing degree dependent on the new internal error tracking system to avoid errors in the workflow.
- A new curation monitoring system: Due to the growth of the number of datasets requesting curation and the existing entries in the Knowledge Graph, the curators are to an increasing degree dependent on an automated curation monitoring system to keep track of the status of datasets in the curation pipeline.
- The Knowledge Graph Search UI has been extended to include multiple new features based on feedback from curators and users (folder download, search by method cluster/modality)
- A curation request evaluation committee has been established
- Nature Scientific Data has included EBRAINS as an approved repository
- Curation of software metadata has been included into the overall curation activities as well as into the ticketing system.
- Integration of previous software catalogue into Knowledge Graph has greatly progressed.

Developments connected to open Metadata Initiative for Neuroscience Data Structures (openMINDS):

- The next generation of openMINDS metadata schemas (v3.0) were developed, implemented and documented on GitHub. The schemas of v3.0 combine the previously used schemas of datasets, models and software that were distributed across v1.0 (MINDS) and v2.0 (uniMINDS).
- The modular structure of the schemas allows users to register their data down to the single file level, extending the findability of data registered to the EBRAINS Knowledge Graph as an openMINDS compatible database.
- The granularity of documenting the techniques, tasks and protocols used to produce the data was improved. Compared to v1.0 and v2.0, the modularity was refined increasing the usage of standardised elements across the database and with that tightening the relations across registered research products (datasets, models, software).
- The schemas were revised to allow the users to define more relations to ontologies, increasing the connections to other data sharing initiatives.

Developments connected to Spatial Anchoring of Neuroscience Data Structures (SANDS)

- The conceptual design of the SANDS metadata schemas has reached maturity and the decision was made to push the further development to the public GitHub as an open source, community-driven project.
- The first version of the SANDS metadata schemas (v1.0) underwent testing for improving the automatised connection between the EBRAINS Knowledge Graph and Interactive Atlas Viewer.
- The upcoming registration of SANDS metadata in the EBRAINS Knowledge Graph (starting in November 2020) will facilitate the access and visualisation of data that were or can be anchored to one of the brain reference atlases supported in the EBRAINS Interactive Atlas Viewer.

Developments connected to In-Depth metadata schemas

- New schemas have been developed to support metadata for EEG, ECoG, MEG, multi-electrode array, calcium imaging, and morphology reconstruction data, and support for structured descriptions of stimulation protocols has been added to all schemas.
- Schemas and a prototype API for tracking provenance in simulation and data analysis workflows have been developed and are being tested by early-adopters.

3.3 EBRAINS Compliance Management, Data Protection, and Data Governance

Over the past half a year, EBRAINS has introduced a number of policies and procedures to assist data users and data providers in finding, sharing and using data on EBRAINS platforms and to ensure that all personal data stored on EBRAINS platforms is secure. This also includes data stored using cloud services and web-based email services.

EBRAINS has also developed a robust data protection and compliance process through the establishment of policies and procedures and most importantly, development of an access review committee that implements the EBRAINS access model of authentication, agreement, and authorization.

These policies and procedures include:

- EBRAINS Access policy[6] and the EBRAINS Access review committee. This system implements the access model of authentication, agreement and authorisation. This is aimed at facilitating the responsible, transparent, extensive and appropriate sharing, access and use of the services and resources provided through EBRAINS. The access policy also ensures that the processes of applying for access to EBRAINS services are simple and clear while facilitating ethically and legally responsible use of research data.
- EBRAINS Data Use Agreement[7] governs the use of pseudonymised human data provided by EBRAINS services to Users and is legally binding.
- EBRAINS General terms of use[8] define the relationship between EBRAINS and individuals that access and use EBRAINS Resources (data, models and software), tools and services.
- EBRAINS Data Provision Protocol (forthcoming). The EBRAINS Data Governance Working group and the Compliance Management, Data Governance and Data Protection team has been working on a protocol to detail the requirements which all data submitted to EBRAINS services should meet. This document is still in development, but will be made available on the EBRAINS website in a short time.
- Integration of ethics compliance procedures into the data curation pathway. EBRAINS compliance management has been working closely with the data curation scientists to ensure that ethics compliance checks are implemented in the data curation pathway in an effective and efficient manner. This should see that data can be processed through any required checks quickly, but in such a way as to ensure compliance with applicable ethical and legal provisions and standards.
- Data Protection Impact Assessment (DPIA). EBRAINS is currently undertaking a DPIA in compliance with Article 35 of the GDPR for services and platforms that interact with EBRAINS research infrastructure. This will detail the nature of data processing, identify risks and develop measures for mitigation of those risks and ensure protection of data subjects.
- The Data Governance Working Group Executive Committee has been developed to oversee critical discussions on maintaining privacy, ethics and responsibility across the Project and to guide the activity of the Data Governance Working Group.

3.4 KnowledgeSpace

In response to community feedback, we have improved the user experience in KnowledgeSpace since April 2020. Specifically:

- The *User Guide* has been significantly overhauled and prominently located on the landing page.

[6] https://ebrains.eu/uploads/EBRAINS_access_policy_v1_1_ed4b84ee9e.pdf
[7] https://ebrains.eu/uploads/EBRAINS_Data_Use_Agreement_90858e7836.pdf
[8] https://ebrains.eu/uploads/EBRAINS_General_Terms_of_use_e457353c1a.pdf

- Filters for the literature panel have been implemented, users can now filter by publication type and journal.
- Accordion style menu has been implemented in the data sources panel. This improvement provides users with a better overview of the types and amount of data available in public repositories related to the entities.
- CURIE and slug are visible in KnowledgeSpace URLs.

In addition, members of the KnowledgeSpace development team have worked closely with the EBRAINS data curation and ontology engineering teams to ensure compatibility and better integration of Knowledge Graph data and models into KnowledgeSpace.

4. New or improved services/functionalities that will become available in the future

4.1 EBRAINS Knowledge Graph

- Before April 2021, a new version of the Knowledge Graph with improved structures will allow users to benefit from simpler access to metadata. It will provide a range of new features (tools, APIs access permission management) and will ensure a consistent migration to the new openMINDS/ SANDS standards. Whilst mostly technical users will benefit directly from the new features in the KG, users of the KG Editor will experience better performance, less temporary inconsistencies (due to a change of the architectural patterns). The users of the KG search will profit from additional information and links between datasets / models and software thanks to the new data structures.
- Additionally, internal rating systems (EBRAINS FAIR rating) will optimise the representation of search results behind the scenes. These will be visually translated into intuitive iconographies that will help users determine if the selected datasets fulfil their requirements by highlighting the best rated ones and the ones most completely described.
- Dissemination efforts will lead to a better spread of information about available and interesting data in the system and will allow users to stay informed about what has been or will be published in the EBRAINS Knowledge Graph.
- Optimisation and extension of the data structures will continue: automated provenance tracking and automated suggestion systems will be added to help enrich the metadata with even more precise information. A continuous optimisation and adaptation to raised user requests will ensure the iterative improvement of all publicly and platform-internal facing user interfaces and therefore ensure the usability over long term.

4.2 EBRAINS Curation

Developments connected to the curation workflow:

- The curation process will add direct interlinks between datasets and the software tools applicable for each dataset: When users search for and find a dataset on the EBRAINS Knowledge Graph Search, they will have the possibility of accessing a tool associated with a particular dataset with interoperable modalities/formats.
- Automated metadata ingestion: Visible to users as a faster curation service.

Developments connected to open Metadata Initiative for Neuroscience Data Structures (openMINDS):

- openMINDS core v3.0 will be integrated as base metadata schemas into the new version of the EBRAINS Knowledge Graph during November 2020. A few exemplary research products registered

with the old openMINDS versions (v1.0 and v2.0) will be migrated to openMINDS. To guarantee a continuous findability for all already registered research products, the migration to the new database system and schemas will be carried out step-wise during 2021.

Developments connected to Spatial Anchoring of Neuroscience Data Structures (SANDS)

- SANDS v1.0 will be integrated as extended metadata schemas into the new version of the EBRAINS Knowledge Graph during November 2020. In parallel, the EBRAINS Interactive Atlas Viewer will establish SANDS as the new interface to gain access to and visualise data that are anchored in one of the supported brain reference atlases. The extension to the SANDS metadata within the EBRAINS Knowledge Graph for relevant research products will be established step-wise during 2021.

Development connected to In-Depth metadata

- Version 2 of the Model Catalogue will be released, including a stand-alone web app, a Collaboratory 2 app, and a REST API. The Model Catalogue is effectively an extension to the Knowledge Graph Search UI that provides additional capabilities specific to models and simulation.
- Similarly, an extension to the Knowledge Graph UI specific to neural activity data will be released, to include visualisation of data analysis and simulation workflows. This will allow the user to see all models and derived datasets that are based on a given experimental dataset, or to follow the pipeline backwards from a figure in a paper through all the analysis steps to the original data.
- In-depth metadata schemas for fMRI data, based on community standards such as BIDS and NIDM, will be developed and used to add in-depth metadata to key fMRI datasets already available in the Knowledge Graph in order to facilitate reuse of these data.
- The schemas for in-depth metadata will be redesigned as formal extensions to openMINDS, reducing redundancy and simplifying metadata access and reuse.

4.3 EBRAINS Compliance Management, Data Protection, and Data Governance

- EBRAINS has developed a Data Provision Protocol, which will assist data providers in sharing their data in EBRAINS platforms. The EBRAINS Data Provision Protocol describes the requirements which data providers must meet to submit data for curation into the Knowledge Graph. This will be made available publicly through the EBRAINS website shortly.
- The EBRAINS Access Policy and Access Review Committee are established. In the first quarter of 2021, EBRAINS will produce a comprehensive guide to the access control model implemented in EBRAINS. This should provide specific guidance to new users who are seeking access to EBRAINS tools and services.
- EBRAINS will shortly publish an Informed Consent Form template that can help guide users in producing their own GDPR-compliant consent documents.
- In the long term, EBRAINS will continue to audit, update and improve processes and policies to ensure that they remain effective, legal and ethically compliant.

4.4 Knowledge Space

- Before April 2021, new data sources, including 2 US Brain Initiative repositories, will be ingested to increase the range of species and data types discoverable through KnowledgeSpace. We also aim to have KnowledgeSpace data and models discoverable through Knowledge Graph search. Release of OpenMINDS and SANDS will enable full integration of Knowledge Graph data and models into the KnowledgeSpace ecosystem.

- Until April 2022, we will continue to increase the range and types of data and models, improve the user experience, ensure tighter compatibility between Knowledge Graph data and models and KnowledgeSpace, and explore how components of KnowledgeSpace (descriptions of neuroscience concepts and literature) can be incorporated into EBRAINS.

Annex 1: User needs

The needs of the users of scientific data and model assets, here referred to as the data and model consumers, have been mapped in a recent study (Research, Nature 2018: State of Open Data 2018. figshare. https://doi.org/10.6084/m9.figshare.7234985.v1).

This study has shown that the large majority of researchers (across all scientific disciplines included) see benefits in having access to others' research data. We have reanalysed data from this study and extracted results from 166 survey participants that are working in the field of biology and medicine and are living in Europe. The results of the reanalysis is shown in Figure 11.

The reported benefits were related to validation of their own findings (28%), complementing existing data (25%), avoiding duplication of efforts (23%), and fostering collaboration (21%). Less than 5% reported that sharing would not benefit them.

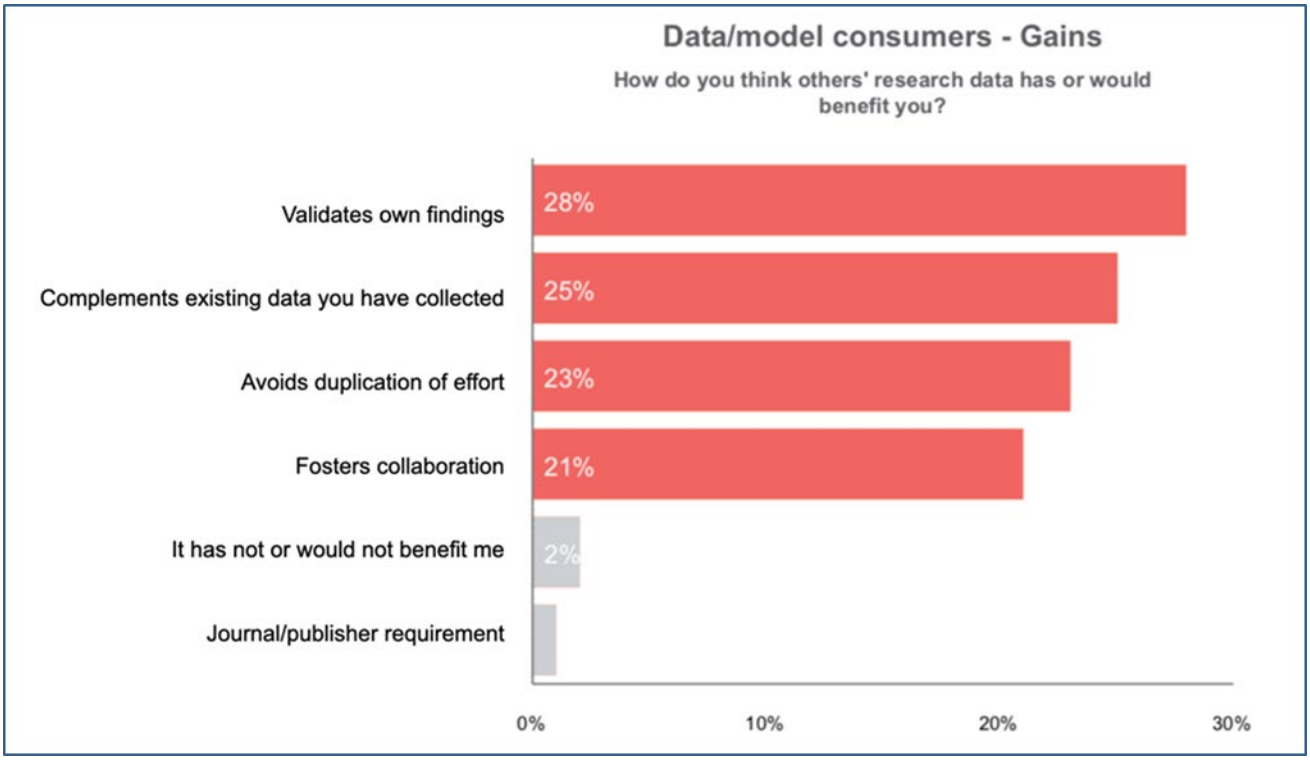

Figure 11: Overview of data and model consumer's view on gains by having access to shared data

Reanalysed from Research, Nature (2018): State of Open Data 2018.

The study also provided an overview of the viewpoints of the data providers. Our reanalysis (same group as reported above) identified the distribution of factors that would motivate sharing of data - the gains for the data providers (Figure 12). Motivating factors included increased impact and visibility of the research, public benefit, and transparency and opportunities for re-use. Other motivating factors were getting proper citation for shared data, and standardisation of sharing data routines. Having trust in the person requesting the data and requirements from institutions or funders were also reported as motivating, as well as general freedom of information.

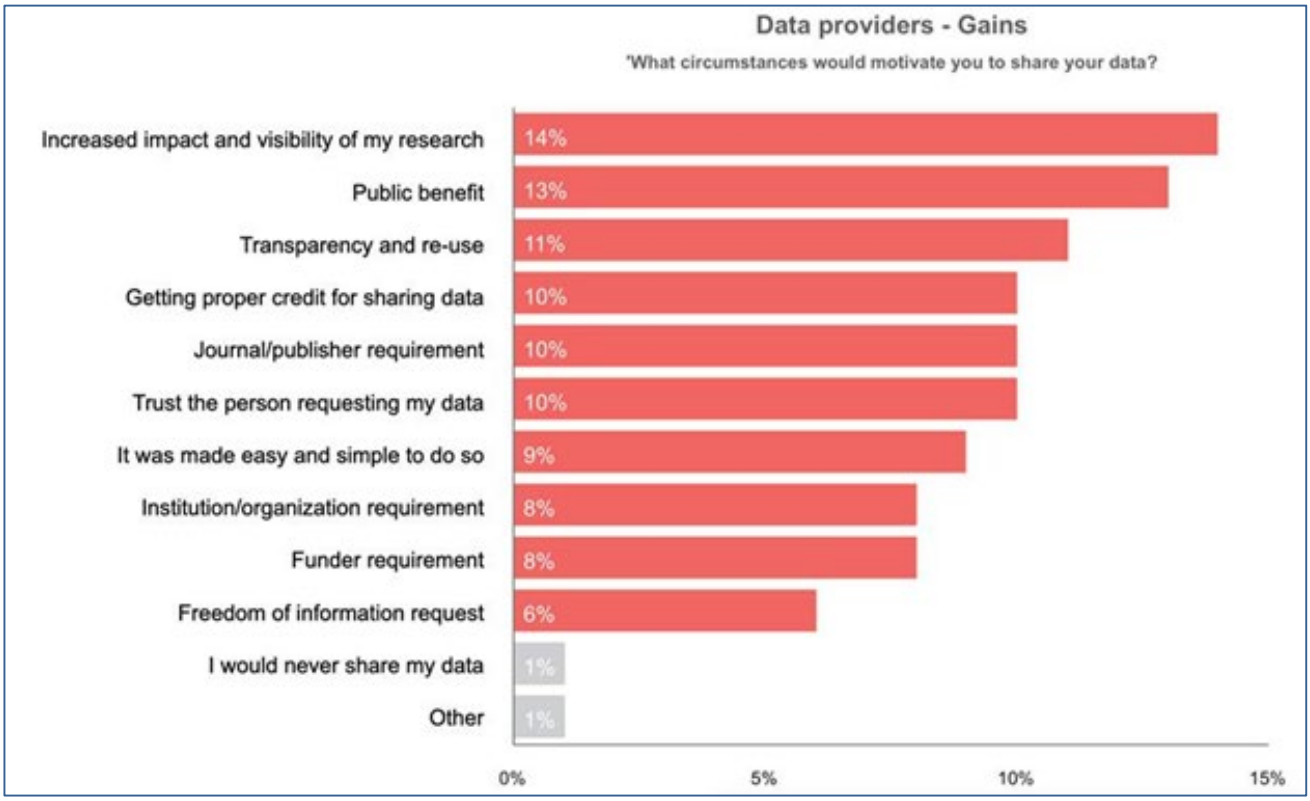

Figure 12: Overview of data providers' view on the gains of sharing data

Reanalysed from Research, Nature (2018): State of Open Data 2018.

Again from the same group, we also identified the distribution factors that were reported as problems or concerns with the sharing of data - the pains for the data providers (Figure 13). The top 5 ranking factors included uncertainties about copyright and licensing, how to organise the data in a presentable form, fear of misuse of data or not receiving appropriate credit, and uncertainty about the right to share.

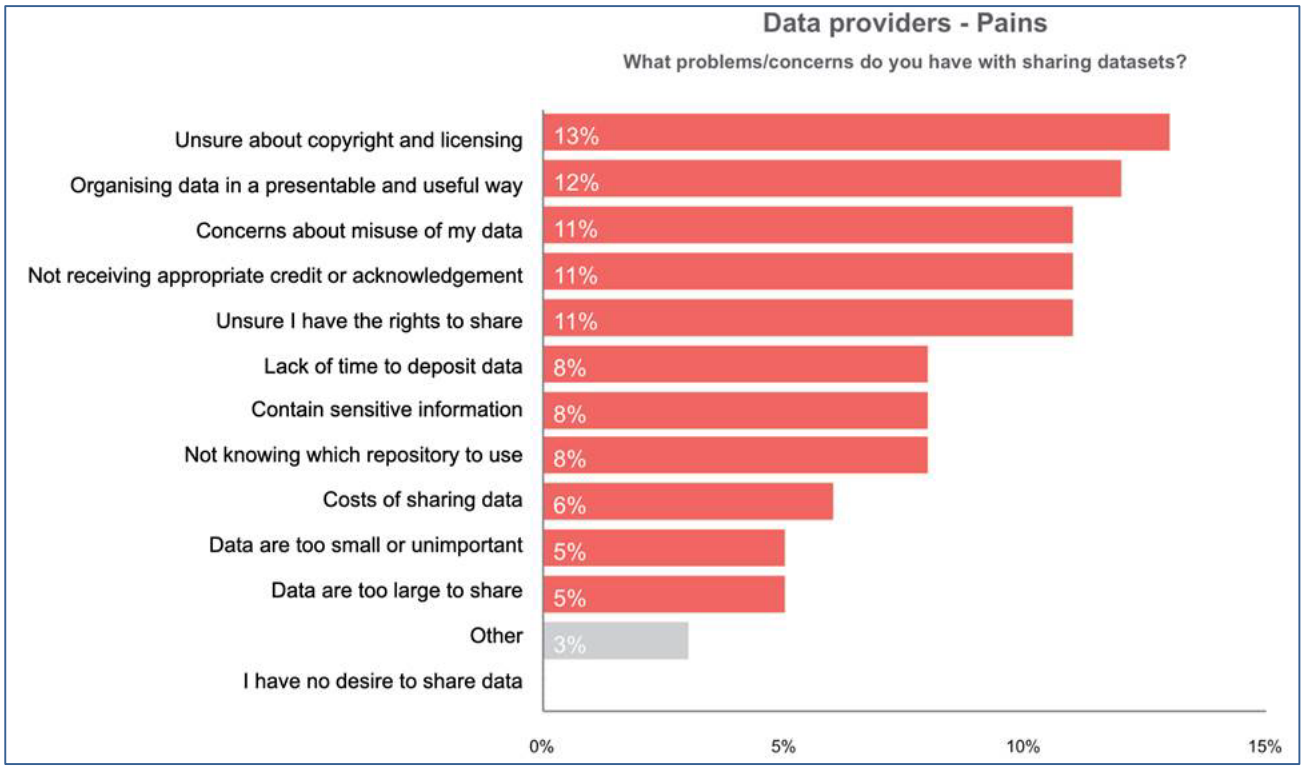

Figure 13: Overview of data providers' view on the "pains" of sharing data

Reanalysed from Research, Nature (2018): State of Open Data 2018.

Other factors reported related to the time spent on preparing data for sharing, sensitivity of the data, not knowing where to share and the practical aspects such as costs and size, and finally whether the data were important enough to share.

Based on the analysis provided above, we have selected 6 documented main user needs for data providers and 6 documented main user needs for data consumers. The EBRAINS Data and Knowledge services are designed to meet these user needs. Figure 14 illustrates how each of the user needs are covered by one or more of the services. For example, the need of the data provider for having their data and metadata organised and data stored is met by services delivered by the Knowledge Graph (1), the basic data and model curation (3), and the Fenix data storage (7). Taken together, we show here how the user requirements translate into the functions provided by the EBRAINS data services.

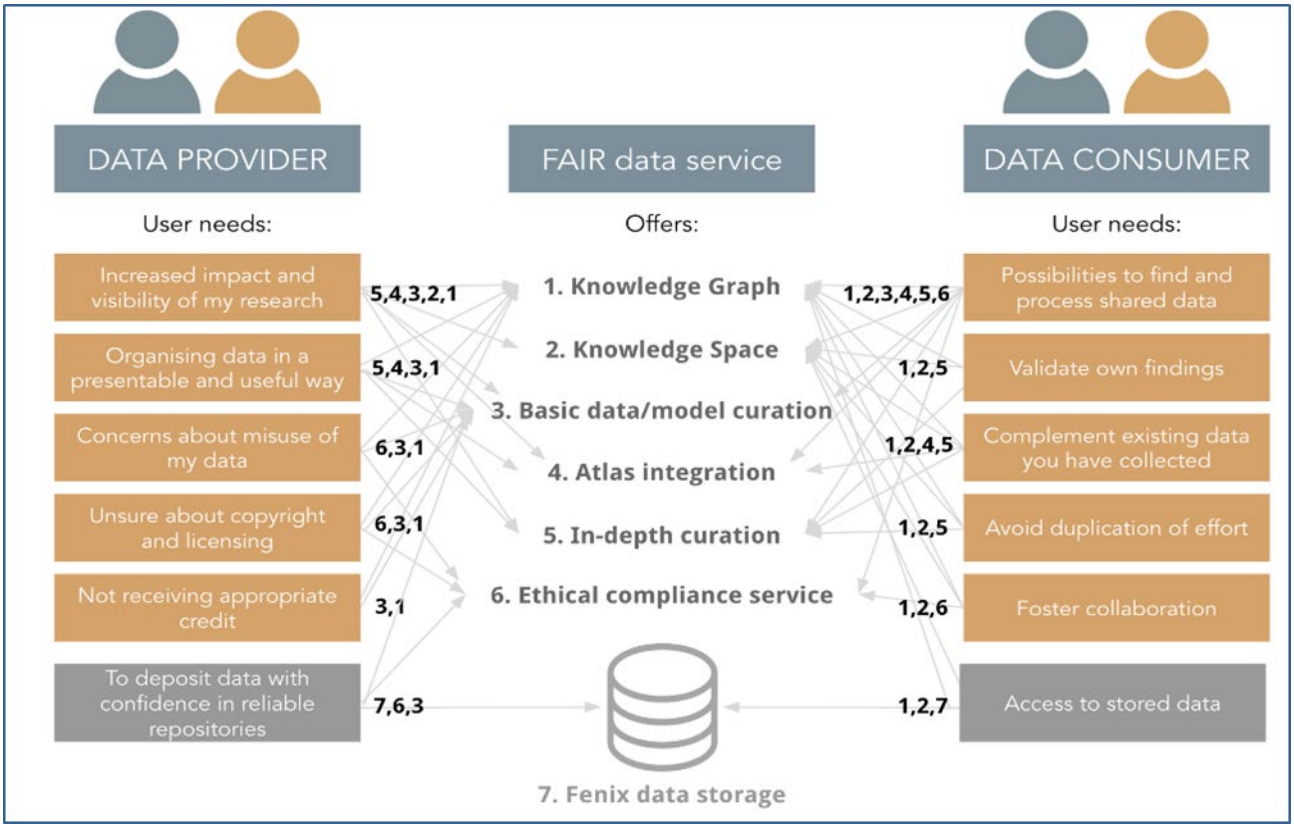

Figure 14: Mapping of user needs documented in Figures 11 - 13 onto EBRAINS services (1 - 7)

Services 1-6 are delivered by the EBRAINS Data and Knowledge services, while service 7 is delivered by EBRAINS Computing services. In combination, the services cover all the major needs of the data providers (left) and data consumers (right). User needs shown are based on Research, Nature (2018): State of Open Data 2018, and High Level Expert Group on Scientific Data. (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data*.

Annex 2: EBRAINS Data and Knowledge services Process Model

The block diagram in Figure 15 shows the processes and steps included in the EBRAINS Data and Knowledge services, the software and hardware components, and the mapping of the user needs onto each step/software/hardware.

The data providers can contribute by registering and uploading data to FENIX storage and metadata to the Knowledge Graph, as well as by adding knowledge to the KnowledgeSpace (of which a subset - e.g. ontologies - are imported as additional metadata as well). The metadata and the uploaded data are enriched, standardised and quality assured by the curation team. The curation process is supported by an automatic inference and metadata management system which detects potential additional relations between metadata blocks, creates derived data (e.g. previews) and notifies about potential missing or invalid metadata.

After the curation process, the data providers review the changes made by the curation team and can either approve or reject them. While a rejection results in another curation iteration, an approval leads to the publication of the data and metadata triggered by the curators. Automatic processes take care of the registration of a DOI, allowing the data to be cited, as well as the registration into the systems used for data and metadata consumption.

Data consumers can find data and knowledge of different granularity either in the KnowledgeSpace or directly in the Knowledge Graph (via the search UI, or by programmatic access via API). If an interesting data structure has been found, the data consumer can further explore the available data by navigating to related datasets, contributors and contact persons and preview and explore the data directly in the Knowledge Graph search results or with the provided viewers.

Once the data consumer has found the data, they can be downloaded (either directly from a web-browser or via a programmatic script) and further analyses can be executed either in individual workflows on external systems or (as recommended) within the Collaboratory, e.g. as part of Jupyter notebooks. The latter allows to easily document, discuss and collaborate on the additional analyses and provides convenience mechanisms to eventually upload and register the derived data and results to the EBRAINS Data and Knowledge services again.

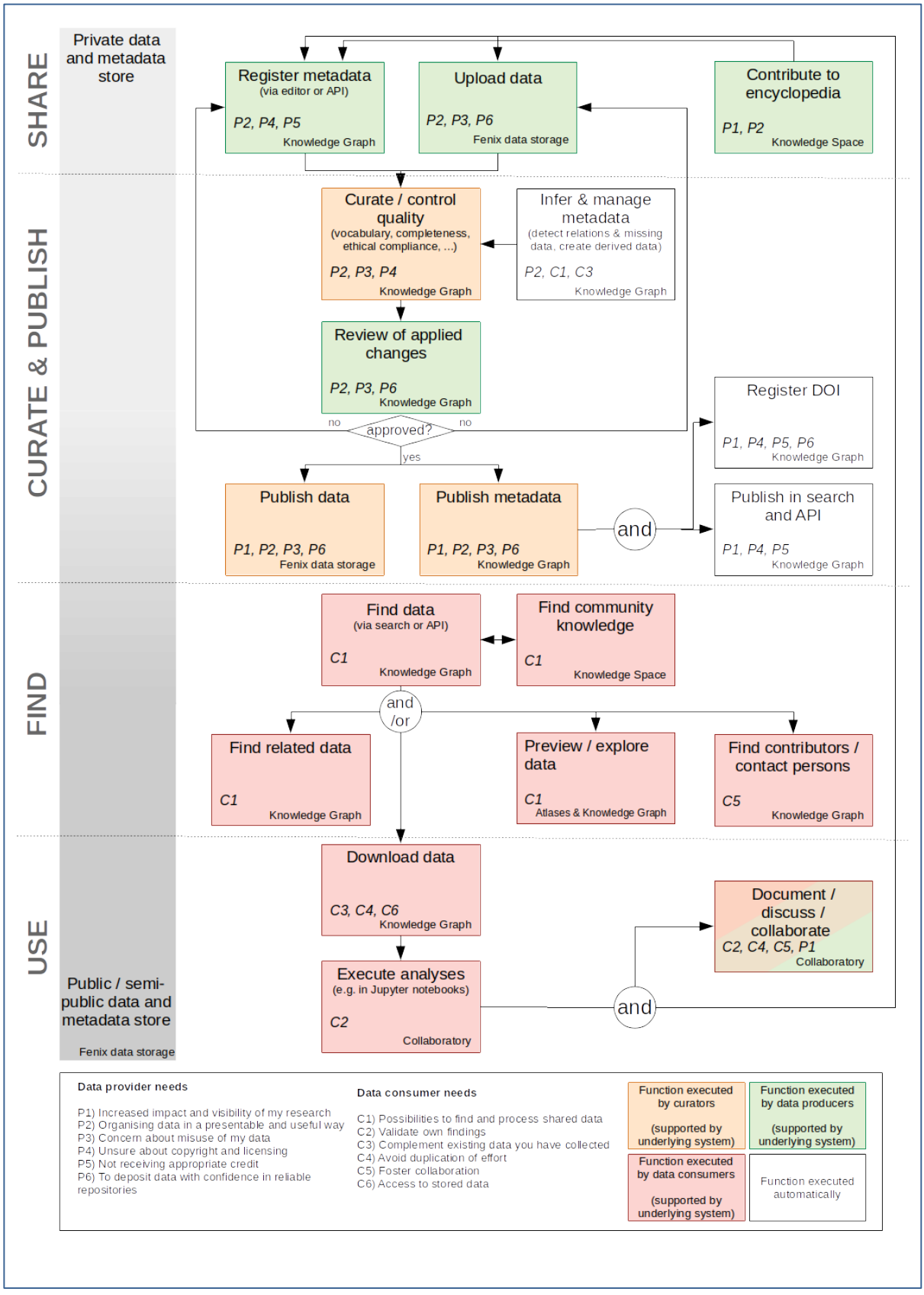

Figure 15: Architecture of the EBRAINS Data and Knowledge services.

Each block in the diagram addresses specific user needs, indicated with P1-P6 for the needs of the data providers (corresponding to user needs shown in the left column of Figure 14), and indicated with C1-C6 for the needs of the data consumers (corresponding to user needs shown in the right column of Figure 14). The blocks are colour-coded according to functions executed by curators (orange), data providers (green), and data consumers (red), and automatically executed functions (white).

Annex 3: Listing of known problems, delays and risks

- Following the release of the openMINDS core, the integration into the EBRAINS Knowledge Graph has started. This integration consists of a number of steps and requires extensive interactions among the Partners delivering the EBRAINS Data and Knowledge services. The full integration is a major effort requiring prioritisation and careful monitoring of the required tasks.
- Several KnowledgeSpace functionalities are depending on developments happening outside of the control of KnowledgeSpace developers. Any delay in the integration of the openMINDS core into the Knowledge Graph will lead to a delay in the integration of data and models from the EBRAINS Knowledge Graph into KnowledgeSpace. Issues or delays in the development of APIs by the US BRAIN Initiative and Brain MINDS would lead to a delay in the indexing of data/models from these initiatives in KnowledgeSpace.
- With a high intensity of efforts focused on developments “under the hood”, and intensive follow up of individual support and curation requests, the much required efforts of the curation team in promoting the overall EBRAINS Data and Knowledge services may suffer. The resources provided by the Inclusive Community Building team and the Outreach team will be utilised in full to deliver targeted and efficient promotion actions.
- The construction of an advanced EBRAINS service for data sharing will have to strike the balance between building the perfect system and creating a user friendly, “good enough”, solution. If this balance is skewed too much toward the perfect system at the cost of user friendliness, potential users may decide to use other services. Communicating the added value of using the service and choosing simplicity when possible in the design of the services will be critical.
- The GDPR gives rise to a number of challenges related to the sharing of human data. EBRAINS Data and Knowledge services have, in close collaboration with the Data Governance Working Group of the HBP, created solutions that define the working space of these services. Close collaboration with the Project’s Communication team will be critical to ensure optimal visibility and clarity of information in this intricate domain.