# Chapter 2

## Getting Started, Getting Data

**Hervé Lemaître, Christopher R. Madan, Declan Quinn, and Robert Whelan**

### Abstract

This chapter explores the availability and accessibility of open-access neuroimaging datasets. It describes how to download datasets using command-line tools (e.g., wget, curl), data management tools such as Datalad, Amazon Web Services (i.e., AWS CLI), and graphical user interface options (e.g., CyberDuck). The chapter emphasizes the importance of accessibility and of documentation for improved research reproducibility. After reading this chapter, researchers will be equipped with the knowledge and tools to download large neuroimaging datasets, including those utilized in this book. We also demonstrate how to download data from OpenNeuro for a range of operating systems.

**Key words** Neuroimaging, Open access, Download, Data management

### 1 Introduction

There has been a remarkable increase in the availability of neuroimaging datasets through open access on the Internet (*see* Madan, 2022 for a comprehensive overview of these datasets, including their significance and diversity [1]). Open-access data allow researchers to conduct neuroimaging studies on over a thousand subjects without the need for scanning them anew. Moreover, this accessibility promotes research reproducibility by enabling the reanalysis of the same data. Notwithstanding the advantages of open-access data, it is important to consider financial or legal agreement issues before downloading these ostensibly "open-access" data. For instance, should the researchers who initially collected the data be included as authors? Should someone coordinate what projects are in progress, in the event that more than one group are working on the same data, and one group might "scoop" another?

One of the pioneering datasets accessible to researchers was the International Consortium for Brain Mapping (ICBM) dataset, which emerged in the late 1990s as a collaborative effort among multiple research institutions [2]. The field of neuroimaging has

Robert Whelan and Hervé Lemaître (eds.), *Methods for Analyzing Large Neuroimaging Datasets*, Neuromethods, vol. 218, https://doi.org/10.1007/978-1-0716-4260-3_2, 

since witnessed substantial growth in initiatives offering open access to data. Some noteworthy examples include the Human Connectome Project (HCP), involving large-scale data collection of many imaging modalities from over 1000 young adults [3], UK Biobank study, aiming to include 100,000 scanned subjects [4], and the ABCD study, which is following more than 10,000 adolescents over 10 years [5].

The nature of the data that can be accessed varies depending on the neuroimaging dataset at hand. For example, data may be either "raw" or "derivative". Raw MRI data require subsequent preprocessing, which can be time intensive (cf. Chapter 16). Conversely, access may be limited to derivative images, eliminating the need for individual preprocessing but also preventing any modification, and limiting control over preprocessing steps. Another distinction concerns individual versus group-level data. Certain platforms provide access to individual subject data, allowing researchers to perform primary analysis at the group level according to their preferences. OpenNeuro (formerly openfMRI) is an example of such a platform [*see* Resources] [6, 7]. On the other hand, some platforms focus on granting access solely to group-level images, facilitating meta-analysis studies. NeuroVault [*see* Resources] is an online platform specifically designed as a repository for sharing, visualizing, and analyzing statistical maps derived from an extensive collection of neuroimaging studies [8].

In this chapter, we will explore various solutions for downloading such datasets, using the AOMIC dataset stored on OpenNeuro as an illustrative example [9]. The complete AOMIC dataset, including all derivatives, occupies approximately 408 GB of storage space. These data can be downloaded via a browser (*see* instructions https://openneuro.org/datasets/ds003097/versions/1.2.1/download); however, this is not recommended for larger datasets if the connection is not stable. Therefore, we demonstrate how to download using a robust method, for Windows, macOS, and Unix. By the end of this tutorial, the reader will be equipped to access other available datasets as well (e.g., HCP, https://www.humanconnectome.org; ADNI, https://adni.loni.usc.edu).

This chapter will focus on importing data to your local machine. It is worth noting that the reverse process also exists. For instance, Coinstac [*see* Resources] is a framework and platform that enables computation to be conducted locally on each participant's machine, while the data remains securely stored at its original source [10]. This approach can be viewed as exporting your analysis without the need to import the actual data, thereby addressing concerns related to data privacy, legal restrictions, and data-sharing agreements.

Throughout this tutorial, command lines will be predominantly employed for Unix-based systems (e.g., Linux, macOS),

specifically Ubuntu, thus requiring basic familiarity with the Unix operating system. If you are using a different Unix operating system, please ensure the availability and installation of the required tools. For those unfamiliar with the Unix operating system, a good explanation of its structure and key components can be found at https://www.javatpoint.com/unix-operating-system. If you are a Windows or Mac user, you can download and use Ubuntu directly (https://ubuntu.com/desktop) for free, or you can try it without committing to major changes to your PC by using a virtual machine (https://ubuntu.com/tutorials/how-to-run-ubuntu-desktop-on-a-virtual-machine-using-virtualbox#1-overview). The following Unix/Ubuntu sections will use the "shell [*see* Glossary]", or command-line/terminal, to download open-source neuroimaging data files, and a good explainer/tutorial can be found at https://ubuntu.com/tutorials/command-line-for-beginners#1-overview. You will also find in the Annex section the specific DataLad instructions for the different chapters of the book.

## 2 Cyberduck (Windows, macOS)

If you prefer a graphical user interface (GUI) for your data transfer needs, there are several tools available (e.g., Filezilla, WinSCP). However, for the purpose of this tutorial, we will specifically use Cyberduck, which is compatible with both macOS and Windows operating systems. At the time of writing, there are no freely available file transfer clients with a GUI for Ubuntu that can establish a connection to the OpenNeuro repository.

Cyberduck [*see* Resources] is a popular file transfer client that supports various protocols, including FTP, SFTP, WebDAV, Amazon S3, and more. It provides a GUI that allows users to connect to different servers and transfer files between their local machine and remote servers. Once you have successfully downloaded and installed the suitable version of Cyberduck for your specific operating system, proceed to launch the Cyberduck application. Locate and click on the "Open Connection" option, as illustrated in Fig. 1, and configure the connection settings as follows:

- Select "Amazon S3"
- In the "Server" field, enter: **s3.amazonaws.com**
- In the "Port" field, enter: **443**
- In the "Access Key ID" field, enter: **anonymous**
- In the "More options" panel and in the "path" field, enter: **/openneuro.org/ds003097/**
- Click on "Connect"

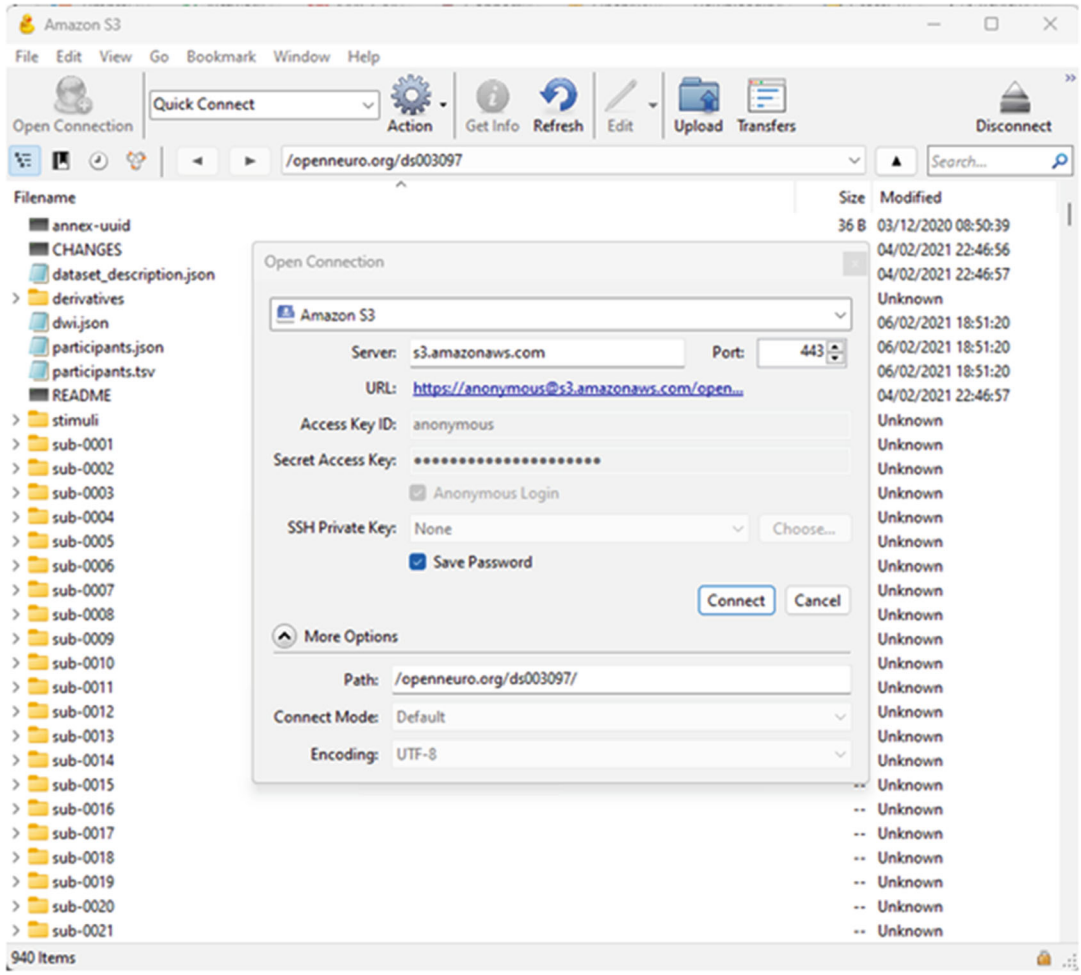


**Fig. 1** Connect to the AOMIC dataset using Cyberduck

Following the aforementioned setup, you should now have the capability to navigate through the entirety of the AOMIC dataset and proceed with downloading ID 1000 data.

## 3 DataLad (Windows, macOS, Unix)

DataLad [*see* Resources] is an open-source data management tool designed to facilitate the management, sharing, and version control of large-scale datasets [11]. The name "DataLad" stands for "Data Lightweight Access and Distribution". It combines the features of data versioning systems, such as Git [*see* Chapter 5], with data distribution capabilities, making it easier to track changes, collaborate on datasets, and ensure reproducibility in scientific research and data analysis workflows.

First, you need to install DataLad on your system (https://www.datalad.org/#install).

In Ubuntu:

```
# Install Datalad
sudo apt-get install datalad
```

**Notes**

- *For Unix, DataLad can also be installed using pip or conda if you are more familiar with these tools.*
- *For Windows, ensure you have Git and Python with pip installed for successful download of DataLad.*

Then, you can use DataLad to download the AOMIC dataset:

```
# Install the AOMIC dataset
datalad install https://github.com/OpenNeuroDatasets/ds003097.git
# Note that this command does not download data per se on your local system but only the data structure
# Download the entire dataset. The get command will actually download and store data on your local system
cd ds003097
datalad get.
# get: actually download and store data on your local system
# Download a subpart of the dataset (the raw data for the first subjects)
cd ds003097
datalad get sub-000*
# You can use the data structure to download any kind of subpart of the dataset
# example:
# sub-0001/anat
# derivatives/freesurfer/sub-0001
```

## 4 AWS (Windows, macOS, Unix)

AWS stands for Amazon Web Services [*see* Resources]. It is a comprehensive cloud computing platform provided by Amazon. AWS offers a wide range of cloud services, including computing power, storage, databases, networking, analytics, machine learning,

artificial intelligence, security, and more. Some neuroimaging datasets are stored on the Amazon Simple Storage Service (S3) for object storage. The AWS CLI (Amazon Web Services Command Line Interface) is a unified command-line tool that can be used to download such neuroimaging datasets.

First, you need to install AWS CLI on your system (https://aws.amazon.com/cli/).

In Ubuntu:

```
# Install awscli
sudo apt-get install awscli
```

Then, you can use AWS CLI to download the AOMIC dataset:

```
# Download the entire AOMIC dataset
aws s3 sync --no-sign-request s3://openneuro.org/ds003097
ds003097

# Download one subfolder (one subject's raw data)
aws s3 cp --no-sign-request s3://openneuro.org/ds003097/sub-
0053 sub-0053 --recursive
# Select and download several subfolders (the raw data for the
first subjects)
aws s3 ls --no-sign-request s3://openneuro.org/ds003097/ --
recursive | \
awk '$NF ~ /^ds003097\/sub-000/ { print $NF }' | \
xargs -I {} aws s3 sync --no-sign-request s3://openneuro.org/
ds003097/{} {}

# the first aws command lists the files
# the awk command filters the lines that matches the pattern
# the xargs command passes the output to the second aws command
for download
```

Alternatively, S3 storage can also be downloaded from using a Python package, Boto[*see* Resources]. This can be useful when you want to selectively download parts of a large dataset such as from the HCP.

## 5 wget and curl (Unix, macOS)

*wget* is a command-line utility for downloading files from the web. It stands for "web get". *wget* allows you to retrieve files from remote servers using various protocols such as HTTP, HTTPS, and FTP. It is a versatile tool that supports recursive downloading, resuming interrupted downloads, following links on web pages, and downloading multiple files simultaneously.

*curl* is a command-line tool and a library for transferring data to or from a server using various protocols, including HTTP, HTTPS, FTP, SFTP, and more. The name "curl" stands for "client URL". With curl, you can send requests to a server and retrieve responses, making it a versatile tool for interacting with web services, downloading files, and performing various network-related tasks.

Tools such as *wget* and *curl* offer the advantage of being readily available on Unix-like operating systems without the need for additional installation. If you wish to download files from the AOMIC dataset on your system using these command lines, you can follow the instructions below:

```
# with wget
wget https://s3.amazonaws.com/openneuro.org/ds003097/sub-0001/anat/sub-0001_run-1_T1w.nii.gz
# with curl
curl -O https://s3.amazonaws.com/openneuro.org/ds003097/sub-0001/anat/sub-0001_run-1_T1w.nii.gz
# -O: saves the downloaded file with the same name as the original file.
```

The method described above only allows for downloading one file at a time, which is not convenient when attempting to download an entire dataset. However, the OpenNeuro website offers a script specifically designed for downloading the complete AOMIC dataset (https://openneuro.org/datasets/ds003097/versions/1.2.1/download). This script navigates through all the files using the curl command.

If you have direct access to the remote directory [*see* Glossary], an alternative option is to employ the wget command for recursive downloads (i.e., to download everything in a folder, including files in subfolders), as curl does not support this functionality.

```
# with wget
wget -r -np http://WEBSITE/DIRECTORY
# -r: enabled recursive retrieval
# -np: avoids ascending to the parent directory when downloading recursively.
```

**Note**
OpenNeuro does not allow recursive access.

## 6 Conclusion

As demonstrated in this chapter, there are various options available for downloading your dataset to your local machine. It is recommended to choose the method that aligns with your operating system and personal experience with either graphical user interfaces (GUIs) or command-line interfaces. It is worth noting that while GUIs generally provide a more user-friendly experience, command-line interfaces offer greater automation potential through scripting. This aspect becomes particularly significant if you intend to re-download the same dataset and document the complete analysis process for your research.

## Annexes

Using Datalad, the subsequent instructions facilitate the retrieval of data for the following chapters to which they are applicable:

***Chapter 4: Establishing a Reproducible and Sustainable Analysis Workflow***

```
# Install the AOMIC dataset
datalad install https://github.com/OpenNeuroDatasets/ds002790.git
# Download the necessary files
cd ds002790
datalad get sub-0001 sub-0002
```

***Chapter 5: Optimizing Your Reproducible Neuroimaging Workflow with Git***

```
# Install the AOMIC dataset
datalad install https://github.com/OpenNeuroDatasets/ds002790.git
# Download the necessary filescd ds002790datalad get derivatives/fs_stats/data-cortical_type-aparc_measure-area_hemi-lh.tsv
```

***Chapter 6: End-to-End Processing of M/EEG Data with BIDS, HED, and EEGLAB***

```
# Install the osf extension for datalad
pip install datalad-osf
# setting up OSF credential as a token (https://osf.io/settings/tokens)
datalad osf-credentials
# Install the OSF repository
datalad install osf://p43rq/
# Download the necessary files
cd p43rq
datalad get *
```

***Chapter 7: Actionable Event Annotation and Analysis in fMRI: A Practical Guide to Event Handling***

```
# Install the osf extension for datalad
pip install datalad-osf
# setting up OSF credential as a token (https://osf.io/settings/tokens)
datalad osf-credentials
# Install the OSF repository
datalad install osf://u5w4j/
# Download the necessary files
cd u5w4j
datalad get *
```

***Chapter 8: Standardized Preprocessing in Neuroimaging: Enhancing Reliability and Reproducibility***

```
# Install the AOMIC dataset
datalad install https://github.com/OpenNeuroDatasets/ds002790.git
# Download the necessary files
cd ds002790
datalad get sub-0021
```

***Chapter 9: Structural MRI and Computational Anatomy***

```
# Clone the AOMIC dataset
datalad clone https://github.com/OpenNeuroDatasets/ds002790.git AOMIC-PIOP2
# Download the necessary files
datalad get -d AOMIC-PIOP2 AOMIC-PIOP2/sub-0111/anat/sub-0111_T1w.nii.gz
# Create an outputs directory and copy the T1w file there
mkdir -p CAT12_derivatives/TEST_sub-0111
cp AOMIC-PIOP2/sub-0111/anat/sub-0111_T1w.nii.gz CAT12_derivatives/TEST_sub-0111/
# delete/drop the local version of the file as we can get it back anytime
datalad drop --what filecontent --reckless kill -d AOMIC-PIOP2 AOMIC-PIOP2/sub-0111
```

***Chapter 10: Diffusion MRI Data Processing and Analysis: A Practical Guide with ExploreDTI***

```
# Install the AOMIC dataset
datalad install https://github.com/OpenNeuroDatasets/ds002790.git
# Download the necessary files
cd ds002790
datalad get sub-*/dwi/
# NICAP data are only accessible through their website
```

***Chapter 13: NBS-Predict: An Easy-to-Use Toolbox for Connectome-Based Machine Learning***

```
# Install the AOMIC dataset
datalad install https://github.com/eminSerin/NBSPredict_-
SpringerNature.git
# Download the necessary files
cd NBSPredict_SpringerNature
datalad get *
```

***Chapter 14: Normative Modeling with the Predictive Clinical Neuroscience Toolkit (PCNtoolkit)***

```
# Install the braincharts data
datalad install https://github.com/predictive-clinical-neu-
roscience/braincharts.git
# Download the necessary files
cd braincharts
datalad get *
```

***Chapter 15: Studying the Connectome at a Large Scale***

```
# Install the AOMIC dataset
datalad install https://github.com/eminSerin/NBSPredict_-
SpringerNature.git
# Download the necessary files
cd NBSPredict_SpringerNature
datalad get *
```

***Chapter 16: Deep Learning Classification Based on Raw MRI Images***

```
# Install the AOMIC dataset
datalad install https://github.com/OpenNeuroDatasets/
ds003097.git
# Download the necessary files
cd ds003097
datalad get participants.tvs
datalad get sub-*/anat/
```

## References


1. Madan CR (2022) Scan once, analyse many: using large open-access neuroimaging datasets to understand the brain. Neuroinformatics 20: 109–137. https://doi.org/10.1007/s12021-021-09519-6
2. Mazziotta JC, Woods R, Iacoboni M, Sicotte N, Yaden K, Tran M, Bean C, Kaplan J, Toga AW, Members of the International Consortium for Brain Mapping (ICBM) (2009) The myth of the normal, average human brain--the ICBM experience: (1) subject screening and eligibility. NeuroImage 44: 914–922. https://doi.org/10.1016/j.neuroimage.2008.07.062
3. Van Essen DC, Smith SM, Barch DM, TEJ B, Yacoub E, Ugurbil K, WU-Minn HCP Consortium (2013) The WU-Minn Human Connectome Project: an overview. NeuroImage 80: 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041
4. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragonu I, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM (2016) Multimodal population brain imaging in the UK Biobank prospective epidemiological

study. Nat Neurosci 19:1523–1536. https://doi.org/10.1038/nn.4393

5. Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, Soules ME, Teslovich T, Dellarco DV, Garavan H, Orr CA, Wager TD, Banich MT, Speer NK, Sutherland MT, Riedel MC, Dick AS, Bjork JM, Thomas KM, Chaarani B, Mejia MH, Hagler DJ, Daniela Cornejo M, Sicat CS, Harms MP, Dosenbach NUF, Rosenberg M, Earl E, Bartsch H, Watts R, Polimeni JR, Kuperman JM, Fair DA, Dale AM, Imaging Acquisition Workgroup ABCD (2018) The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. Dev Cogn Neurosci 32:43–54. https://doi.org/10.1016/j.dcn.2018.03.001
6. Poldrack RA, Gorgolewski KJ (2017) OpenfMRI: open sharing of task fMRI data. NeuroImage 144:259–261. https://doi.org/10.1016/j.neuroimage.2015.05.073
7. Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, Hardcastle N, Wexler J, Esteban O, Goncavles M, Jwa A, Poldrack R (2021) The OpenNeuro resource for sharing of neuroscience data. eLife 10:e71774. https://doi.org/10.7554/eLife.71774
8. Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, Sochat VV, Nichols TE, Poldrack RA, Poline J-B, Yarkoni T, Margulies DS (2015) NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front Neuroinformatics 9(8). https://doi.org/10.3389/fninf.2015.00008
9. Snoek L, van der Miesen MM, Beemsterboer T, van der Leij A, Eigenhuis A, Steven Scholte H (2021) The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. Sci Data 8:85. https://doi.org/10.1038/s41597-021-00870-6
10. Plis SM, Sarwate AD, Wood D, Dieringer C, Landis D, Reed C, Panta SR, Turner JA, Shoemaker JM, Carter KW, Thompson P, Hutchison K, Calhoun VD (2016) COINSTAC: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. Front Neurosci 10:365. https://doi.org/10.3389/fnins.2016.00365
11. Halchenko YO, Meyer K, Poldrack B, Solanky DS, Wagner AS, Gors J, MacFarlane D, Pustina D, Sochat V, Ghosh SS, Mönch C, Markiewicz CJ, Waite L, Shlyakhter I, de la Vega A, Hayashi S, Häusler CO, Poline J-B, Kadelka T, Skytén K, Jarecka D, Kennedy D, Strauss T, Cieslak M, Vavra P, Ioanas H-I, Schneider R, Pflüger M, Haxby JV, Eickhoff SB, Hanke M (2021) DataLad: distributed system for joint management of code, data, and their relationship. J Open Source Softw 6:3262. https://doi.org/10.21105/joss.03262