README Files for Datasets

A Guide for Writing README Files for Datasets

The ASAP Open Science Team
Dana Lewis, PhD
Matthew Lewis, PhD
Devin Snyder, PhD
Robert Thibault, PhD

README Files for Datasets

The purpose of this document is to provide guidance on how to write README files to accompany datasets.

The ASAP Open Science Policy requires all research outputs, including datasets, be deposited in a publicly accessible repository no later than the time of publication and cited in the publication with a permanent identifier. **ASAP requires a README file to be included in the upload of all datasets.**

To assist ASAP grantees in sharing usable data and to ensure that they meet the ASAP data-sharing standards outlined in this document, **the Open Science Team will review all ASAP-affiliated Zenodo dataset uploads for inclusion of a README file.**

If you have additional questions, email the Open Science Team at **openscience@parkinsonsroadmap.org** with the subject title: "Readme for Datasets Question:...".

Table of Contents
- What is metadata?
- What is a readme file?
- How do I make a readme file?
- What should be included in a readme file?
- Readme Best Practices
- Why does ASAP require a readme file for datasets?

What is metadata?

Metadata is information about the data, which may include how it was collected, units of measurement, how files relate to each other, and a description of how to use the data. Metadata should be specific enough that it differentiates a dataset from any other, similar dataset.

Specific disciplines and datatypes have guidance or standards for what metadata should be shared with data, some of which can be found in the ASAP Data Repository Wiki.

What is a readme file?

A **README** file contains the metadata of an associated dataset. A readme file should provide a clear, concise description of relevant details about data collection, processing, and analysis. The goal of a readme file is for anyone, now or in the future, to be able to understand the structure and content of the datasets and reuse them.

How do I make a readme file?

A README file can be created using a plain text editor, like Windows NotePad or Mac TextEdit, or using MarkDown. We recommend using this ASAP template to make your own README file.

What should be included in a readme file?

Your README file should include the following items:

1. **General Information**
    - Include the date the README file was generated. If you update the file, you should also indicate that here.
    - This section should include a brief description of the dataset, including the date(s) of data collection and any software dependencies.

> **REMINDER:** Software dependencies are any software necessary to use or view a dataset.

2. File Overview
- File naming conventions (how files are named)
- List of file formats included in the dataset
- Relationship between the files or a description of the file structure

> **TIP:** Use a systematic file naming convention that conveys important experimental metadata.

3. Data Specific Information
- This section is data file specific. For each file, you should include:
    - A brief description of the data in each file
    - List and define column headings in tabular data, units of measurement, date formats, variables, acronyms, abbreviations, labels, scoring of categorical data.
    - Explain how missing data are represented (e.g., NA or blank).

4. Methodology
- Instruments and software used to collect the data
- Workflow and processing steps to replicate analysis; this includes scripts, code, packages, notebooks, and software used to process the data

> **TIP:** You should only share methodology that might be relevant if someone wants to (a) combine your dataset with others for meta-analysis or (b) re-analyze your dataset.

5. Data Access and Sharing
- Publications based on this dataset
- Recommended citation for this dataset
- Datasets should be licensed with a CC BY 4.0 or CC0 license.
- You should acknowledge ASAP in your data upload using the following language:

> *"This research was funded in whole or in part by Aligning Science Across Parkinson's [ASAP Grant #] through the Michael J. Fox Foundation for Parkinson's Research (MJFF)."*

README Best Practices

File and Folder Organization
- Metadata should be recorded in the repository itself *and* as a readme file included in the upload.
- Use an underscore in the name of the README file (e.g., _README) so it appears at the top of the file structure.
- A README file should be accessible from the same location as the dataset and saved at the root folder/directory of the dataset.
  - For complex datasets, you may want to provide multiple README files. However, there should still be a main README file that contains basic information about the project as a whole.

Writing and Language Conventions
- Write as clearly as possible; don't use jargon or acronyms.
- Write for broad understanding and reusability. Don't assume anything is common knowledge.

Why does ASAP recommend a readme file for datasets?

README files have many benefits, including:
- Helping others understand the contents, provenance, how to interact with and interpret your dataset
- Making data understandable and reusable for you and for others you may with to re-use your dataset
- Creating and updating a README file during active research helps you track your work and makes it easier to share your data when it is ready to publish
- Ensuring long-term preservation and re-use of potential datasets

Aligning Science
Across Parkinson's