AGEL DMP Overview

Last Updated: 2025.11.11 by C. Watson

One part of research that most researchers hate – admin tasks. Researchers should try to make their research products available to the AGEL shared repository during ongoing research. After publication, researchers should additionally archive the published form of their data in an online public repository (e.g. Harvard Dataverse, Dryad, or OSF) that is (ideally) linked in the published manuscript.

The goal of this document is to ensure that all AGEL researchers are adhering to a common standard to ensure consistency and ease of access across all AGEL data, no matter who worked on it. This guarantees that all AGEL-related research products (raw/reduced data, codes, etc.) follow F.A.I.R. data principles.

This document will cover:
- AGEL Data Policies
  - This covers topics such as guidelines for file naming conventions and data sharing policies

- Archiving Data in the AGEL Dataverse
  - This covers setting up your dataverse account, creating your own dataverse, and linking to the main AGEL dataverse

- Making a README file
  - This section has lots of links to resources to help in creating README files that should *always* be included in any data that is uploaded to the dataverse

AGEL Data Policies

Last Updated: 2025.11.11 by C. Watson

Here are some guidelines and suggestions of file naming conventions and policies on data sharing in regards to AGEL data.

File Structures and Naming Conventions

Whether you are working with imaging, spectra, or modeling data, here are some general guidelines of how to organize and name your data to facilitate data sharing later on (i.e., so that when we dig up your research in 5 years we know exactly what the file AGEL000645-442950A_F200LP_WFC3_level2.fits contains and how it was made without having to track you down to interrogate you). Part of the naming convention will include a flag of the data product level. These are defined on the next page.

- ➢ **Imaging**
    - ○ <u>file format(s)</u>: FITS
        - ■ Imaging files should only be in FITS formatting
    - ○ <u>file structure:</u> target / instrument / proposalid_filter
        - ■ Including the proposal id number can be helpful for distinguishing between older and new observations (particularly for HST imaging where we have multiple follow ups)
    - ○ <u>naming conventions</u>: agelname_filter_camera/instrument_[reduction level].fits
        - ■ (e.g., a drizzled science image from HST may be named 'AGEL000645-442950A_F200LP_WFC3_level2.fits')
    - ○ <u>required metadata:</u> included readme describing reduction process in addition to filling out metadata in AGEL Dataverse
    - ○ <u>required ancillary files</u>: reduced images + things that were integral to the reduction process (check with PI)

- ➢ **Spectra**
    - ○ <u>file format(s):</u> FITS, ASCII
        - ■ ASCII files should only be released if it's a basic format of two columns, wavelength and flux
    - ○ <u>file structure:</u> target / instrument / proposalid
    - ○ <u>naming conventions:</u> agelname_instrument_[instrument specific info]_[reduction level].fits
    - ○ <u>required metadata:</u> readme with description of data and reduction process in addition to filling out metadata in AGEL Dataverse
    - ○ <u>required files:</u> reduced spectral file

➢ **Models**
  ○ <u>file format(s):</u> HDF5
  ○ <u>file structure:</u>
  ○ <u>naming conventions</u>:
  ○ <u>required metadata</u>: readme with description of data, modeling software used, and best-fit results in addition to filling out metadata in AGEL Dataverse

➢ **Figures/Tables/Other Data**
  ○ <u>file format(s)</u>: EPS, PNG, JPG, PDF, ASCII
    ■ Figures should be saved in eps or png format preferably (pdf = big file sizes usually so should be avoided if possible)
    ■ Table data should be saved as plain text file in ASCII format, with header information at top of file and all columns have names
  ○ <u>naming conventions</u>: AGEL_{project/user specific info}_[reduction level]
    ■ Since, by the time you are creating figures and data tables, you are likely using higher-level data products, your figures/tables/other data will likely fall under a 'level 3' product, which should be reflected in the file naming.
  ○ <u>required metadata:</u> readme file explaining what is contained within the figure/table. For tables, the readme should include a breakdown of the columns and their units. For figures, include any line fits, if applicable, and include references to points that come from other papers. Also include a short description of how the asset was created – did you use matplotlib to plot the figure or maybe astropy to build a table? What conditions were placed, what cuts were made, etc.

Data Product Levels

As mentioned in the previous section, part of your file naming convention should include a flag of the data product. This is a simple integer that will denote what level the data is at:

Level #		Notes	Examples
Level 0	Has not been modified by researchers	These will not likely end up in your final data release – unless you have HST imaging, in which case you should release the raw versions too	Raw data from telescope, archive data, etc.

Level 1	Reduced single run observation	Combined exposures from one visit but *not* combined mosaics) This level will represent the majority of the data in the public AGEL data releases – calibrated and ready for others to do science	Imaging has been calibrated (flat fielded, CR rejected) and aligned to common WCS frame – e.g., the AGEL HST data in filters F140W, F200LP, and F606W Spectra have been wavelength calibrated, sky subtracted (or equivalent), etc. – e.g, AGEL DR2 spectra
Level 2	High level science data	Deep combined data/mosaics – the products you are more likely to be doing hard science with	Imaging is a combined moasic comprising multiple visits and/or filters – e.g., HST lookbooks, color images, etc. Spectra are combined SEDs
Level 3	Community contributed high level science products	Specific to your project but not general AGEL DR#	Model fits, spectral fits, figures, tables, etc.

Level 1 products will encompass the majority of the curated public AGEL data releases – essentially the barebones data that we have at least calibrated and gotten to a science-ready version that can be taken by other researchers to make their level 2 and 3 products.

By simply adding this flag to the end of your filenames (e.g. as '*_level2.fits'), you can easily (1) help your future self when you inevitably forget what is actually in the file and (2) help fellow AGEL researchers so that when you share a level 2+ product, we can have the confidence of knowing that the data was handled a certain way and can be used for science.

Data Sharing Policies

As stated in the AGEL Project Policy document, "Unless specifically stated, all observations collected are freely available to team members. However, AGEL data is not to be shared beyond the collaboration except by request and approval by the collaboration. When in doubt about whether data or results can be shared beyond the AGEL collaboration, ask the leadership team."
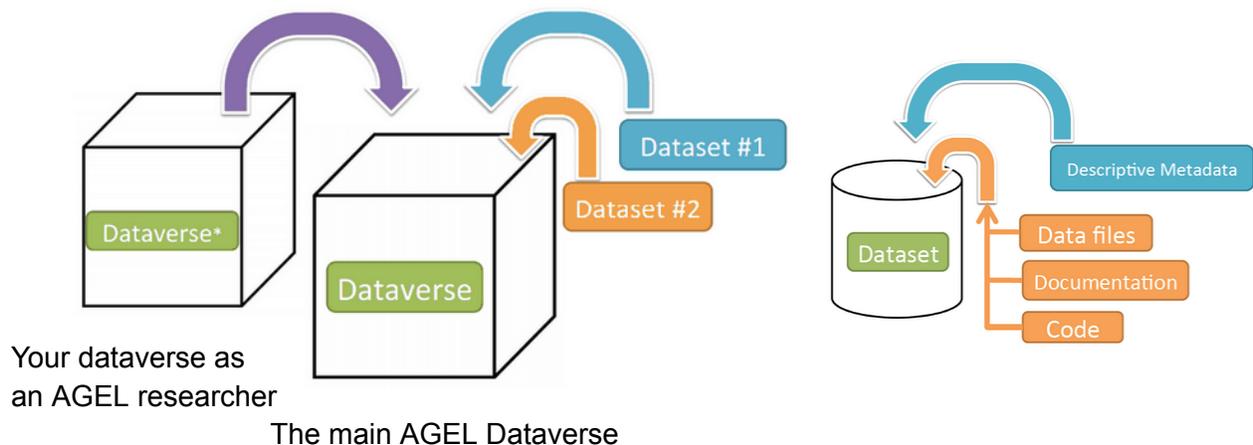
The AGEL Dataverse

Last Updated: 2025.11.06 by C. Watson

The majority of the publicly available and curated AGEL data should be published online at the AGEL Dataverse, an online repository for storing your project files (similar to GitHub but with more metadata and ability to create/link DOI) – e.g., observational data (raw and reduced), catalog tables, scripts used for analysis, etc. Instructions for Admins to add users to certain dataverses is included at the end of here.

What is a Dataverse vs Dataset?

The main AGEL Dataverse collection will house all public data release datasets and individual researchers' dataverses which contain data from the AGEL survey. A **dataset** will contain the actual data files and documentation associated with those files and will be saved within either your individual dataverse (that is linked to AGEL) or the main AGEL Dataverse.

Your dataverse as
an AGEL researcher

The main AGEL Dataverse

The main AGEL Dataverse currently houses two sub-dataverses:
- Published work using AGEL Data
  - This will house all datasets associated with individual researcher's papers/projects
  - Anyone with a Dataverse account should be able to upload data to this dataverse
- Observational Data
  - This will house all curated public level 1 data products, organized by instrument/filter/wavelength of observation

- - - Users will need to be given access in order to be able to upload to this dataverse
  - [AGEL Public Data Releases](#)
    - This will house all curated public data releases (e.g., those packaged as DR#)
    - Users will need to be given access in order to be able to upload to this dataverse

---

Creating your Dataverse account and your first dataset

- **Create account** on [Harvard Dataverse](#) (doesn't matter what institution you are from, free to make account)
  - can link your ORCID (very helpful for security reasons as well so we can make sure that fully registered users can curate the AGEL Dataverse)

- **Organize & document your data** into a single parent folder (doesn't matter what this folder is named, once you upload to website, only subfolders remain viewable – so make sure they are named in a way that makes sense)

  - Include a README file either for each subdirectory you are uploading or (preferably) one README file in the parent directory that explains everything: what the project is or what the data represents, file formats, column headings, units when necessary, etc.

  - Best practice for README is to follow [CDS standards for Vizier tables](#)

  - Also helpful to use [AAS online converter tool to create machine readable tables](#)

- **Upload Data**. You can either
  - (1) Create a Dataverse where you can save data, files, scripts, etc. -- could be useful to create one Dataverse for your project so that you can easily link the generated data DOIs to your paper when publishing

  - (2) Create a Dataset - Main package of your data + documentation – **Note: you can upload whole directories *after* creating the dataset** (but readability can be an issue if you don't **toggle the view to "Tree" mode**)

  - ★ Whatever you chose, **make sure to change** the 'Host Dataverse' to either the main [AGEL Dataverse](#) (for general data releases) or '[Published work using AGEL data](#)'

- Once you have created your Dataverse and/or Dataset, make sure to **populate as much metadata as possible**. Required fields for linking AGEL data are:
  - Title

- ○ Author
- ○ Point of Contact (who should be reached in the future for questions regarding this data and reduction?)
- ○ Description -- be as descriptive as possible here or include a README with additional info -- what is the data? What format is it in? What project was it used for? Is it raw or reduced? Are the reduction scripts available in your Dataverse too? Etc.
- ○ Subject -- set to Astronomy and Astrophysics
- ○ Note: After you create your dataverse/dataset, you can edit the metadata or change access permissions.

- ● NOTE: You do not have to publish your dataverse or dataset right away, you can instead generate a draft link. If you hover your mouse over 'Edit Dataset' a drop-down menu will pop-up with an option to generate a preview url. This way you can send your PI or data czar a preview to make sure that everything looks good for publishing.
  - ○ Alternatively, it can be good to get your data uploaded (but not yet published) to include a citable DOI to your own data in a submitted paper. Then when the paper is accepted you can publish the dataset.
  - ○ This would also give your reviewer private access to view your dataset if you include your preview link when submitting your paper

- ● Once you are ready for your data to be made public, make sure to **Publish your Dataverse and/or dataset**. This will allow the auto-generated DOI to be made searchable and publicly available.

The Dataverse limit is 1TB *per researcher* (2.5TB for those with harvard or affiliate emails) on datasets, with a 2.5GB file size limit on uploads. So individual researchers can upload their data as their own Dataverse/dataset that is then linked to the main AGEL dataverse, limiting the main uploads to the AGEL Dataverse to just the curated public data releases. If we need to go beyond this, the Harvard Dataverse people may be willing to work with us (I have been told that they are itching for more astronomy data), but for large datasets (2.5TB+), they may charge an estimated $11/TB/year for data storage.

➡ Best practice: Create a dataverse or dataset containing your data (following the AGEL guidelines defined in this document). Link your dataverse/dataset to the main AGEL Dataverse (Upload research products to the central AGEL repository (following the defined AGEL standards) and add links to your data in the Data Roadmap.

Dataverse Admins

Users who are designated as 'admin' role have the ability to give new users access to the observations and data release dataverses by going to the dataverse page for which access needs to be given, click on 'Edit'→ 'Permissions'→ click on 'Users/Groups' → click on 'Assign Roles to Users/Groups' → search for name of who you need to add (if they do not show up, then they have not created a dataverse account yet and will need to do that first) → select the role to assign them (usually Contributor is best, or Curator if you reeeeallllyyyyy trust them)

AGEL Data README Files

Last Updated: 2025.10.23 by C. Watson

It is very important that you develop the habit early on in your career of creating README files (this is going to be another one of those things that your mentors never follow and kick themselves for, so they try really hard to get you to develop these habits early to make your lives easier than theirs)

- Helpful to use [AAS online converter tool to create machine readable tables](#) (MRTs) to generate MRTs in correct formatting (and this generates header information that can then be copied into a more general README for the project.

- Also helpful to review the [AAS Data Guide](#)

- Best practice for README is to follow [CDS standards for Vizier tables](#) and [also here](#)

At minimum, your README should include

- Your name, affiliation, contact information, and the date that you are uploading the data (make sure to update if you had uploaded to dataverse earlier and just saved it as a draft before publishing later!)
- Citation to paper if associated with specific one
- Name of dataset (either title of paper or AGEL DR#)
- List of top contributors (likely author list of paper)
- Description of project and data (what were you trying to do and what data did you use/collect to do the thing?)
- Summary of directories and/or files
- For tables, include a description of each column in the table, what the column headings are (aka your parameter names), what the units are, and short descriptions of what it is ([see example here](#))
- Policy for data use and sharing – e.g. something that says this data is being made publicly available but ask them to cite your paper or this dataset (via the dataverse DOI) and include AGEL survey acknowledgement (from [AGEL Project Policy](#))

There are also python tools (e.g., [cdspyreadme](#)) that can be used to generate CDS formatted README files.