Survey Data Support Services at CSDC / Astro Data Lab

Requirements & Recommendations

Robert Nikutta, Mike Fitzpatrick, Joanna Thomas-Osip, Oliver Oberdorf, Dara Norman
Version: 2023-07-25

Table of Contents

1. Synopsis

The purpose of this document is to outline the data services available to NOIRLab Survey proposals and Gemini Large and Long Programs (LLPs) for their datasets through the Astro Data Lab[1] science platform (Data Lab for short), to define the requirements and make recommendations regarding dataset organization, data formats, and metadata content, with the goal of maximizing the usefulness of survey data products to the entire astronomical community, and to make the process easy for both the PIs and CSDC/Data Lab.

The document also outlines expectations regarding timelines and communication channels.

2. Available services at Data Lab

NOIRLab's Astro Data Lab science platform (SP) provides several capabilities which fall broadly into two categories: hosting & access to survey-level large datasets, and providing co-located compute and data services around the data. We briefly outline the capabilities and services below.

2.1. JupyterLab notebook / compute environment

Data Lab runs JupyterLab notebook servers for its users, co-located with the data holdings at Data Lab and at NOIRLab's Astro Data Archive[2]. The co-location of large data and compute resources enables access to and analysis of very large datasets without the need to download any data, nor to install any software. The only things needed are a Data Lab account (free), an internet connection, and a good idea.

Note that data products can be downloaded (for instance using the `datalab` command-line client tool) anonymously, without a Data Lab account. A Data Lab account is needed when the Data Lab Jupyter notebook server, the MyDB personal database, and the VOSpace remote file storage capabilities are to be used.

Finally, note that the compute environment also includes the user's resources (for instance uploaded tables, files, etc.), which can be seamlessly used together with the survey datasets hosted at Data Lab.

2.2. Catalogs (TAP)

Tabular data (catalogs) are ingested into Postgres databases and can be queried using Data Lab's own `queryClient` (both in a Jupyter notebook environment and through the `datalab` command-line client), through Data Lab's web query form, and through any other TAP[3]-aware clients such as TOPCAT[4]/STILTS[5].

[1] Astro Data Lab: https://datalab.noirlab.edu
[2] Astro Data Archive: https://astroarchive.noirlab.edu/
[3] TAP - Table Access Protocol, a Virtual Observatory protocol for accessing remote catalog data.
[4] TOPCAT: http://www.star.bris.ac.uk/~mbt/topcat/
[5] STILTS: http://www.star.bris.ac.uk/~mbt/stilts/

2.3. Simple Cone Search (SCS)

A simple cone search service can be stood up for catalogs containing positional information. From the IVOA description of SCS: "[SCS] defines a simple query protocol for retrieving records from a catalog of astronomical sources. The query describes sky position and an angular distance, defining a cone on the sky. The response returns a list of astronomical sources from the catalog whose positions lie within the cone, formatted as a VOTable."

Note that at DL the same functionality can be achieved using the Q3C[6] extension for Postgres (installed at DL), specifically the `q3c_radial_query()` function.

2.4. Crossmatch tables

Data Lab can host crossmatch tables. Crossmatch tables computed by the Survey/LLP teams are also possible and welcome. If the Survey/LLP data products comprise multiple tables, we highly recommend that a common and unique key (column) be used between them where appropriate (for instance an `objectid` column or similar).

If a Survey/LLP computes crossmatches between several of their own tables, or between their own tables and external datasets, the method/mode of creation should be clearly stated to avoid confusion (e.g., single nearest neighbor match, or based on Bayesian probability density, etc.)

By default Data Lab computes crossmatch tables of an ingested object table against reference astrometric, photometric, and spectroscopic tables. Currently these are:

Table 2.4.1: Reference datasets used for crossmatch tables at Data Lab (as of mid 2023)

Reference table	**Kind**
gaia_dr3.gaia_source	astrometric
allwise.source	photometric
nsc_dr2.object	photometric
unwise_dr1.object	photometric
sdss_dr17.specobj	spectroscopic

The crossmatch tables computed by Data Lab:

- are only computed if there is footprint overlap
- use a default matching radius of 1.5 arcseconds
- only keep the single-nearest neighbor for each source
- do not keep rows without matches in the other table
- two crossmatch tables are computed for each reference table: object_table-to-reference_table and reference_table-to-object_table. The two tables are not identical, in general.

If the Survey/LLP team thinks that other crossmatch tables against datasets hosted at Data Lab would be useful to the scientific community, this should be discussed with Data Lab upfront and

[6] Q3C: https://github.com/segasai/q3c

early in the process. The currently hosted survey tables, including crossmatch tables, are listed in Data Lab schema browser.[7]

2.5. Crossmatch service

Data Lab provides a fast crossmatch capability through a web tool[8] , and cross-matches are also possible in Jupyter notebooks and Python scripts (via DL libraries). Users can upload or generate their own data tables, and crossmatch them with any table hosted at Data Lab.

Users can (currently) not perform crossmatches between two tables hosted by Data Lab using the web-based tool, but many such tables are already pre-computed. Note that such crossmatches are possible using the Python libraries.

2.6. Simple Image Access (SIA) and Image Cutout

The IVOA Simple Image Access protocol (SIA) provides capabilities for the discovery, description, access, and retrieval of image datasets.

At Data Lab, image collections (as FITS files) are run through a metadata scraper, and the information extracted from the headers populate a separate database which allows SIA queries (for image discovery). SIA queries allow filtering the list of images based on the positional overlap on the sky, filter band, exposure time, data product, and many more characteristics. Certain FITS header keywords are required for SIA during the metadata scraping, and several others are highly recommended. Please see 3.2.3 for details.

Image cutouts can be requested for specified positions, including keywords for the FOV and others. SIA query results contain fields with URLs for image download and image cutouts.

2.7. File Service

A heterogeneous collection of survey files, with an arbitrary directory structure, can be served through Data Lab's File Service. Typically, the File Service directory structure is exactly as the Data Release directory structure that Data Lab receives from a survey team. The benefit is that access to any file can follow the same path mechanism that may be already familiar to the survey team members and/or software.

Each survey is assigned a named handle that exposes the files in its File Service to public access via Data Lab's `storeClient` in, e.g., a notebook environment, and the `datalab` command-line utility.

Example: `gogreen_dr1://` is the File Service handle/prefix for the GOGREEN/GCLASS LLP Data Release 1.

Named file services can be listed/discovered within Data Lab.

2.8. Spectroscopic data

Fast access to spectroscopic data is currently being developed at Data Lab as the SPARCL[9] service. While the initial spectroscopic surveys included are SDSS and DESI, Data Lab intends to support arbitrary spectroscopic data products in the future. Until then, spectra (based on FITS files) can be made accessible through the standard Data Lab File Service, and read using standard tools.

We intend to retroactively ingest past LLP spectroscopic datasets into the new service sometime in

[7] Browse the currently hosted catalogs at Data Lab: https://datalab.noirlab.edu/query.php
[8] Web-based crossmatch service: https://datalab.noirlab.edu/xmatch.php
[9] SPectra Analysis and Retrievable Catalog Lab: https://astrosparcl.datalab.noirlab.edu

the future. That is why we ask that the necessary metadata (keywords in FITS files) be provided already now (see Sections For Images and Spectra and For Spectra).

2.9. Survey/LLP-specific example and use-case Jupyter notebooks

Jupyter notebooks provided by the Survey/LLP teams are very helpful to users, and we highly recommend that the Survey/LLP teams provide at least one notebook (but multiple notebooks are welcome, each, for instance, showcasing a different aspect or use-case of the dataset).

The Data Lab team can help in developing such notebooks or converting them to work within the Data Lab system. Example notebooks can range from demonstrating various data access modes to the Survey/LLP data products at Data Lab, to data reduction steps, and science use cases with the LLP data.

Notebooks become part of the Data Lab notebook suite that every Data Lab user has access to. They are being developed and hosted on GitHub[10] and are made available immediately to all Data Lab users through a link within their JupyterLab notebook environments.

2.10. Survey landing page (website)

Each Survey/LLP is represented on the Data Lab website through a survey landing page. This should serve as the first entry point for users new to the Survey/LLP data products and can introduce the program, summarize the data products, and link to data release papers, etc.

Data Lab provides the Survey/LLP team with a simple landing page template that should be used as a starting point and should be filled with relevant information by the Survey/LLP team.

For inspiration, please see some of the already existing landing pages.[11]

3. Requirements and recommendations

3.1. General

The Survey/LLP data products should be generally "publication ready". Data Lab won't reorganize files or do internal crossmatches of measurement tables to make the data useful outside the original survey. The Survey/LLP team will need to decide how the data release products should be organized to be scientifically valuable to others.

If the Survey/LLP data products comprise a large collection, we might try a sample ingest and standing up of services before doing a bulk transfer. This will allow us to estimate the total ingest time required, and the full data would need to be delivered to us at least that many days/weeks prior to an agreed date for release. We want to avoid, e.g., a holiday loading rush before the winter AAS meeting etc.

The Survey/LLP should also provide the exact number of rows and columns for every table. This is for “zero-order” quality assurance.

3.2. Files, data formats, and metadata

3.2.1. For catalogs

- FITS bin tables are strongly preferred over other formats (e.g. CSV)

[10] Data Lab example notebook collection: https://github.com/astro-datalab/notebooks-latest/
[11] Survey landing pages at Data Lab: https://datalab.noirlab.edu/survey.php

- All files must pass `fitsverify` cleanly
- Table descriptions are needed for all tables
- Column descriptions are needed for all columns
- Column names must only contain small-caps ASCII characters, the numbers 0-9, and the underscore '_' character. The first character must not be a number. The column name must not contain any SQL reserved keyword. The column name must be at most 59 characters long.
- Column descriptions should be provided in separate CSV files (ideally one CSV file per table).
- Ensure that data type representation in FITS files does not do unexpected things to the data (e.g. loss of accuracy due to too small dtypes, signed vs unsigned dtypes, etc.)
- Unique identifiers for each object are required, and should be carried over to other tables that may be used for lookup (e.g. object vs measurement tables); ID data type can be a number of things, e.g., INT, BIGINT, CHAR.
- JD/MJD column present for measurement tables
- Units wherever possible or where values are possibly ambiguous (e.g. flux, photometric system, errors)
- Primary positions should be in decimal degrees (J2000)

If a Survey/LLP has already internally created a database, Data Lab can also accept Postgres SQL dump files of the DB (which would retain the schema).

3.2.2. For file service

- Any heterogeneous collection of files relevant to the survey
- Directory structure created with files sorted into directories
- Can also hold things like JPG/PNG files, config files, misc files
- All "detritus" files have been removed (no Trash folder, no `.DS_Store`, no tmp files, etc.)

3.2.3. For both images and spectra

In general, any FITS file containing image data, spectra, or binary tables, must satisfy these requirements:

- Simple FITS or MEF are acceptable.
- They must pass `fitsverify` cleanly
- Flux units must be present.

Some FITS header keywords mentioned below apply generally, and are either required, or highly recommended. Others may apply specifically to Gemini data products.

The following keywords are required. If they are Gemini-specific (eg. GEMPRGID) then they are required if the files are from a Gemini instrument. They allow ingestion into Gemini Observatory Archive (GOA) and also potentially help with the association of raw products. Some of these keywords are also necessary for the SIA and image cutout capability at Data Lab.

- TELESCOP - This records which site the observation is from (Gemini-North or Gemini-South)

- INSTRUME - The instrument used for the observation
- OBJECT - The target object
- DATE-OBS - The time of the observation
- GEMPRGID - This is the Program ID used at Gemini and will relate this data to any other related observations within that semester.
- PROPID - The proposal ID under which the data were collected
- OBSID - This is the observation ID
- OBSMODE - 'IMAGE' for images, 'SPECT' or 'LS' for spectroscopy, 'MOS' for multi-image spectroscopy and 'IFU' for integral field spectroscopy.
- FILTER1, FILTER2 - The filter wheel setting
- EXPTIME - The exposure time in seconds (if applicable)

The following keywords also have significant value but are less important than the previous list. They are recommended, but not required.

- SEQEXVER - This is the version of software that was used for observing
- OBSERVER - This was the observer who ran the observation
- SSA - This is the System Support Associate on duty at the time of the observation
- DISPERSR - grating/prism for spectroscopic modes
- CWAVE - central wavelength for spectroscopic modes
- FPMASK - The focal plane mask
- XBIN, YBIN - The X and Y binning, in the extension with the binned pixel data
- HA - Telescope hour angle (i.e. '+2:25:26.05')
- AIRMASS - Mean airmass for the observation

3.2.4. For images only

- WCS is required, image center and corners are helpful
- All positions in decimal degrees
- Seeing estimate and magnitude zero point, if available
- Pipeline/software name and version used for reduction and analysis

3.2.5. For spectra only

- Should contain POS/BAND/TIME conceptual values for data discovery, and the exact keywords used should be indicated to the Data Lab team
- SNR/redshift/resolution/object classifications are helpful
- Flux and wavelength calibration levels
- Since there is no "standard" spectra format, FITS tables or image extensions are both okay
- Log-linear dispersion is preferred, but not required
- Observed vs rest frame must be stated
- In the future: extracted 1-D spectra can be served through the SPARCL spectroscopic service at Data Lab; for a list of SPARCL core fields, please see: https://astrosparcl.datalab.noirlab.edu/sparc/sfc/
  Original echelle/IFU/MOS/etc. spectra can only be served through the Data Lab File Service.
- Object ID should tie to the associated catalog table when possible
- Prefer to have flux/model/sky and ivar or error arrays

- Any line list would be treated as catalog dat

3.3. Expectations regarding communications and responsibilities

The Survey/LLP team will designate one main point of contact (POC) within the collaboration to be working closely with the Data Lab team. This can be the PI or someone designated by the Survey/LLP team.

Data Lab will also designate one main POC to communicate with the Survey/LLP POC.

Delegation of communication is of course possible when the collaboration between Survey/LLP POC and Data Lab has been well established.

3.4. Timelines

The time to ingest Survey/LLP data products and to stand up data services around them can vary strongly depending on data volume, the time of the year (e.g. before AAS meetings), and on the currently available cycles at Data Lab. Therefore, early communication with Survey/LLP Teams that are preparing data products for ingestion is highly recommended and appreciated by the Data Lab project.

The POCs shall first establish the expected date for production readiness, and work backward to the necessary timeline and milestones for data delivery, any possible test ingestions, full ingestion, and the production of a survey landing page and example Jupyter notebooks.

If a Survey/LLP release shall coincide with, for instance, a data release paper or a conference/workshop, it is even more important that communication begins early. Three to six months is a realistic minimum to ensure a smooth process. An even longer lead time is appreciated.

3.5. Gemini Provenance Considerations (LLPs only)

These are additional suggestions for any FITS files to get the most value out of the Gemini Observatory Archive. There are some header keywords and extensions that have special meaning to Gemini. When data are reduced using Gemini IRAF or DRAGONS[12], the keywords will be managed automatically. If these can be maintained in the final FITS file data products, it will make it easier to connect to raw and calibration files.

3.5.1. PROVENANCE Extensions

From DRAGONS v3 onward, reduced data will have provenance information recording the reductions performed. These are stored in two table extensions: ‘PROVENANCE’ and ‘PROVENANCE_HISTORY’. If you are using DRAGONS, these can be carried forward and preserved in any delivered FITS files. This can be used to link back to the source data (raw and calibrations). This is optional, but highly recommended if available.

3.5.2. IRAF Provenance Keywords

For IRAF reduced data, these keywords can be used to track the provenance of data products.

ORIGIN - The IRAF Kernel used

IRAF-TLM - Time of last modification

BIASIM - File used for bias correction

[12] https://www.gemini.edu/observing/phase-iii/reducing-data/dragons-data-reduction-software

GSFLATIM - File used for flat correction

3.5.3. Provenance Mapping File

We can also support a provenance mapping file for those who do not use DRAGONS for their data reduction. This will allow files to be related back to their source data and is not limited to FITS files. This should be a simple CSV file where each line has a delivered filename followed by one or more source datafiles. For example:

```
science_file1.pdf, rawfile1.fits, rawfile2.fits
science_file2.fits, rawfile2.fits, rawfile3.fits
```

or alternatively

```
science_file1.pdf, rawfile1.fits
science_file1.pdf, rawfile2.fits
science_file2.fits, rawfile2.fits
science_file2.fits, rawfile3.fits
```

4. Glossary

AAS	American Astronomical Society
CSDC	Community Science and Data Center
DB	Database
DL	Data Lab, short for “Astro Data Lab science platform”
File Service	File-level access to arbitrary survey files; collection of all the files that make up a Survey/LLP data release
GOA	Gemini Observatory Archive
HLSP	High-level science product; data product(s) produced by a Survey or LLP
LLP	A single Gemini Long and Large Program (has LLP ID and at least one PI)
PI	Principal Investigator of a Survey/LLP program
POC	Point of contact (on either Survey/LLP or Data Lab side)
SCS	IVOA Simple Cone Search
SIA	IVOA Simple Image Access
SP	Science platform
SPARCL	SPectra Analysis and Retrievable Catalog Lab
TAP	IVOA Table Access Protocol (for catalogs in databases)