NOIRLab Astro Data Archive

Sean McManus & Knut Olsen (NOIRLab)

An archive of raw, reduced, and high-level data products is critical to maximizing the scientific productivity of an observatory. Archives act as multipliers of scientific impact and accessibility and prepare their communities to become frontline users of observatory facilities. The NOIRLab Community Science and Data Center (CSDC) maintains a multi-petabyte-scale library spanning 25+ years of data taken at Cerro Tololo Inter-American Observatory (CTIO) and Kitt Peak National Observatory (KPNO). This resource comprises data acquired through the National Optical Astronomy Observatory (NOAO) and partner (e.g., WIYN, SOAR, SMARTS) observing programs. As NOAO merges into the new NOIRLab organization, CSDC continues to maintain stewardship of this essential asset for the astronomical community. To that end, sustained investments are needed in computer hardware and software technologies to ensure continued access going forward.

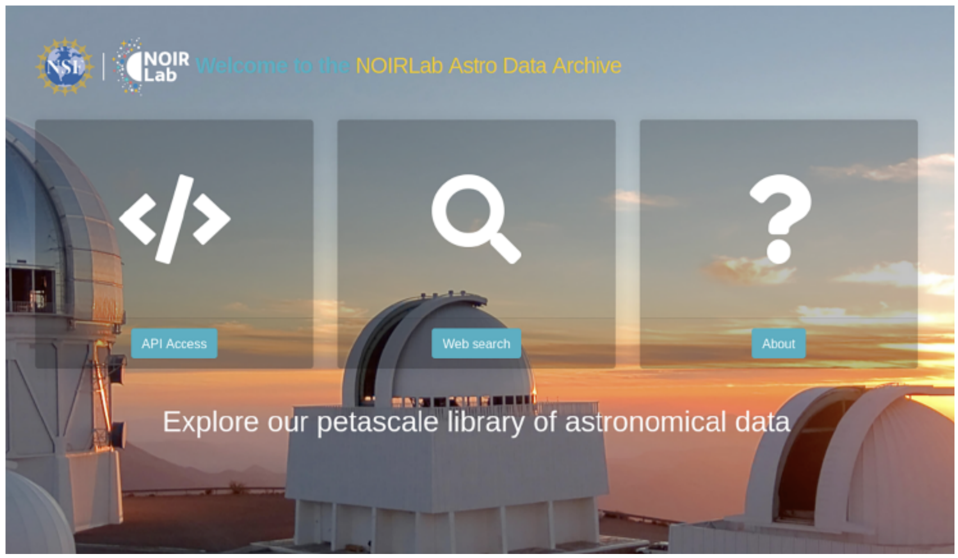

Figure 1. NOIRLab Astro Data Archive home page. (Credit: NOIRLab/NSF/AURA)

To meet the growing technological needs of data-intensive astronomy, CSDC has rolled out a new data archive platform, the NOIRLab Astro Data Archive (Figure 1).

The new software system is a complete replacement of the legacy NOAO Science Archive and surpasses its predecessor in capabilities and maintainability. The Astro Data Archive continues the core mission of providing long-term access to public and proprietary telescope data products derived from PI and survey-scale observing programs. In addition, new tools offer enhanced capabilities that add value to data that might have otherwise have been left "dark" or not easily discoverable.

The new Archive source code is built on open-source software, including Django, Postgres, Vue.js, and Elasticsearch. Leveraging these libraries has shrunk the codebase to 20% of that of the legacy system. With the smaller, more nimble software stack, CSDC software engineers are able to quickly deliver new services to users.

While serving a long legacy of historic astronomy data, the Astro Data Archive is also on the front line of capturing data from ongoing operations at CTIO and KPNO, including the hundreds of gigabytes produced each night with the Dark Energy Camera (DECam; mainly funded by the US Department of Energy) on the Víctor M. Blanco 4m Telescope. Images captured with DECam and other instruments are typically stored in FITS (Flexible Image Transport System). FITS images include both data (pixels) and metadata, which consists of ancillary information related to an observation, such as proposal, principal investigator (PI), epoch, filter, and target information.

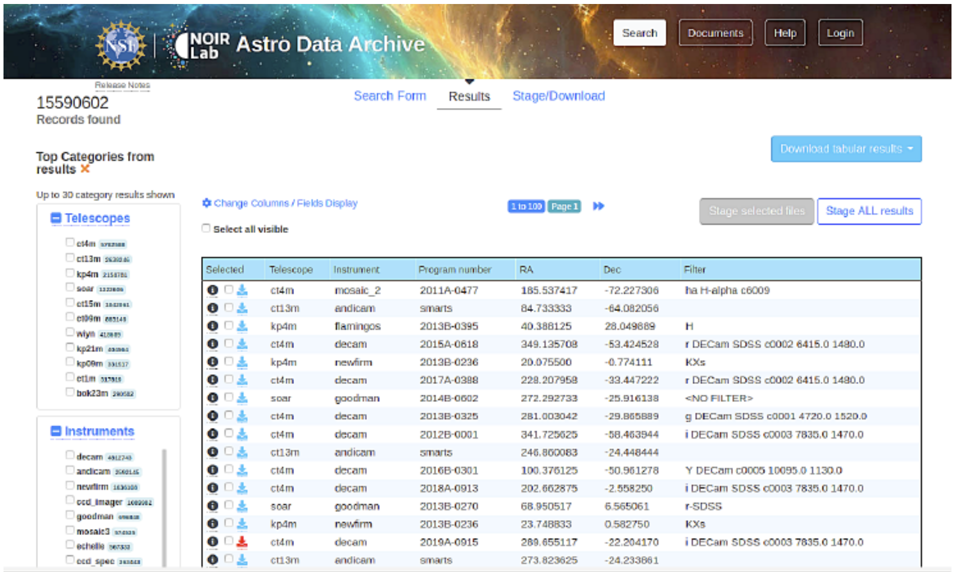

Figure 2. Astro Data Archive search results page and category filters. Users can filter and sort by common fields, view image headers, and download images or tabular data. (Credit: NOIRLab/NSF/AURA)

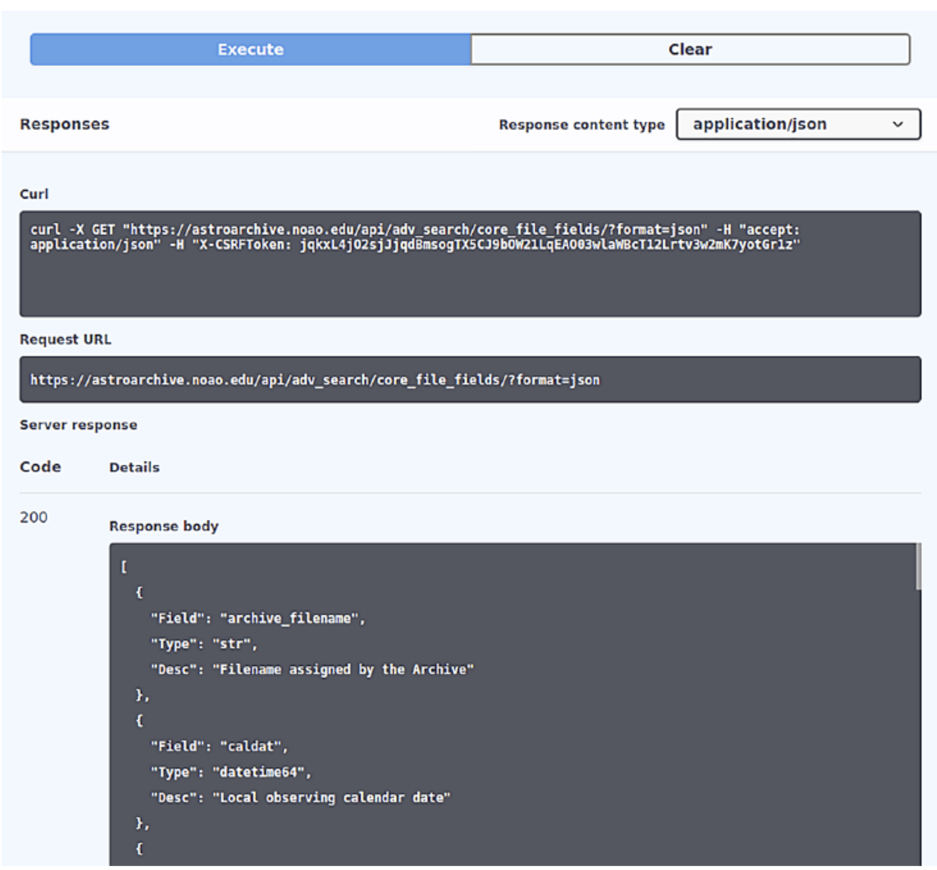

Figure 3. Screenshot of Astro Data Archive live documentation. This page describes basic API usage and options with real working examples that users can edit and try out. In this particular example, the core file fields service provides common field descriptions available for search for all instruments. (Credit: NOIRLab/NSF/AURA)

Because the Archive stores millions of files from nearly 40 different custom telescope and instrument combinations spanning decades, it is a challenge to normalize heterogeneous data sufficiently to make them discoverable to a simple user interface. The new system abstracts this complication into so-called “personalities,” which are applied when the data are ingested. Personalities define a set of instructions on how to process metadata for each instrument and provide the mapping functions needed to translate raw FITS keywords into commonly known search fields (e.g., RA, DEC, FILTER). Some instrument-specific keywords don’t map to well-understood fields but are still potentially valuable to the end user (e.g., engineering data). Those extra keywords are also captured into a special data type that is searchable via the application program interface (API) Advanced Search mechanism (see below).

A main goal of the Astro Data Archive is to offer several modes of data access and to make petabytes of file holdings transparent to a variety of use cases for users with different interests and technical expertise. This is accomplished by providing complementary graphical and programmatic search interfaces.

The graphical user interface aims to provide a convenient means of data discovery and distribution to the user, whether it be a PI seeking their observing run or a general researcher looking for information on a particular target. Similar to most search engines, users are able to provide as much or as little information as desired to explore the entire petabyte-scale Archive. The graphical user interface is backed by Elasticsearch, which is an open-source distributed search engine. It works similarly to Google in the sense that search results are delivered nearly instantaneously by a scaled back end, which facilitates ad hoc data exploration and discovery. Users can add or remove filters to their search without providing a priori information (Figure 2).

To meet the demands of data-intensive science, a new API is available for researchers seeking programmatic entry to the Archive. The API offers many service endpoints for varying use cases. For example, the well-established Simple Image Access protocol is available for quick data discovery with RA/DEC position and radius. In addition, the Advanced Search service offers an interface for deep searches across the entire Archive, including instrument-specific metadata. API search results are available in many formats (CSV, JSON, VOTable) and deliver direct access to image-pixel data through a download service.

To lower the learning curve of using programmatic interfaces, the Archive API endpoints are described and illustrated through live documentation (Figure 3) that includes working examples of search specifications and results in common formats. In addition, science workflows are demonstrated through Python Jupyter notebooks, and the NOIRLab module is provided through Astroquery.

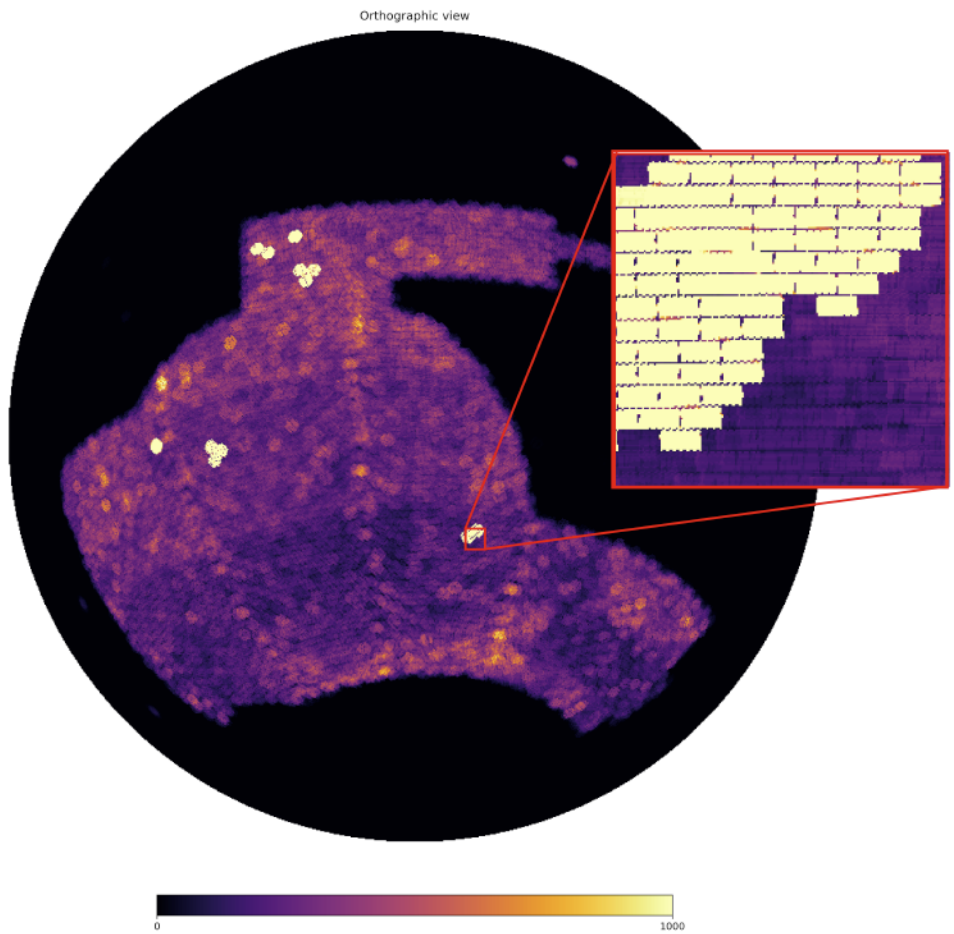

Figure 4. Effective exposure time for the Dark Energy Survey, as computed by the exposure-map.ipynb notebook found in the Archive GitHub repository. The larger map shows the effective exposure time variation across the DES footprint, which is largely uniform except for features introduced by the deep supernova fields, seams in the tiling scheme, and variations in observing conditions. The inset, which features a small region near a pair of supernova fields, shows the map's resolution, which is finer than that of the individual Dark Energy Camera CCDs.
(Credit: NOIRLab/NSF/AURA)

As an example use case of the Astro Data Archive's new capabilities, a user wanting to probe the Dark Energy Survey (DES) DR1 catalog (served through the NOIRLab Astro Data Lab) for new Milky Way structures would likely need the spatially dependent DES depth map to uniformly measure the density of detected objects. This depth depends on the exposure time of all of the images covering each location in the survey footprint and the seeing, sky brightness, and transparency when those images were taken. As demonstrated in the exposure-map.ipynb notebook linked from the Astro Data Archive, the archive is able to retrieve, within a few minutes, information on transparency and exposure time from the headers of 50,000+ full-focal-plane DECam images in seconds and information on image quality, sky brightness, and CCD corner coordinates for the more than 3 million individual DECam CCDs contained in those images. With this information in hand, the notebook shows how to create a high-resolution map of total effective exposure time, shown in Figure 4.

Visit the Astro Data Archive website to give the interface a try, and contact us if you have ideas for more example use cases that we could showcase in notebooks!