

# PROCEEDINGS OF SPIE

[SPIDigitalLibrary.org/conference-proceedings-of-spie](https://spiedigitallibrary.org/conference-proceedings-of-spie)

## The NASA/IPAC Infrared Science Archive (IRSA) as a resource in supporting observatory operations

G. Bruce Berriman

G. Bruce Berriman, "The NASA/IPAC Infrared Science Archive (IRSA) as a resource in supporting observatory operations," Proc. SPIE 7016, Observatory Operations: Strategies, Processes, and Systems II, 701618 (12 July 2008); doi: 10.1117/12.788635

**SPIE.**

Event: SPIE Astronomical Telescopes + Instrumentation, 2008, Marseille, France

# The NASA/IPAC Infrared Science Archive (IRSA) as a Resource in Supporting Observatory Operations

G. Bruce Berriman<sup>1,a</sup>

<sup>a</sup> Infrared Processing and Analysis Center, 100-22 California Institute of Technology, Pasadena, CA, USA 9112

## ABSTRACT

IRSA's scaleable and extensible architecture is inherited by new missions and data providers, and thus offers substantial cost savings to missions. It has built archives for the W.M. Keck Observatory & the Spitzer Space Telescope Legacy teams, among others. It provided archiving and databases support for 2MASS, when active, and will provide corresponding support for the forthcoming WISE mission. IRSA acts as a resource to projects and missions by advising on product design and providing tools for validating data products.

**Keywords:** Archives, infrared astronomy, data products, ground-based observatories, space missions

## 1. INTRODUCTION

The NASA/Infrared Processing and Analysis Center (IPAC) Infrared Science Archive (IRSA) is the archive node for NASA's infrared and sub-millimeter astronomy projects and missions. IRSA (<http://irsa.ipac.caltech.edu>) opened for business in 1999, to curate and serve data from the Infrared Astronomy Satellite (IRAS) mission and to provide product generation and database management support for the Two Micron All-Sky Survey (2MASS), which was then beginning operations. The multi-Terabyte scale of the 2MASS data products led IRSA to develop a software architecture designed to provide real-time access to very large data sets. Today, IRSA serves data from 24 projects and missions and its science holdings are 23 TB in size.

IRSA is chartered to provide three functions:

- Serve and curate NASA's infrared and sub-millimeter data sets, associated documentation and ancillary products;
- Enable optimal scientific exploitation of these data sets by astronomers
- Support planning for, and operation of, new missions and projects.

The third item of IRSA's charter will be the subject of this paper. IRSA provides support to observatories and missions in the following areas:

- Deploying archives on behalf of observatories and missions
- Support observation planning
- Support pipeline processing
- Support the production of new data sets
- Support preparations for new missions and projects
- Support observing proposal submission

---

<sup>1</sup> [gb@ipac.caltech.edu](mailto:gb@ipac.caltech.edu) ; phone 1 626 395-1817; fax 1 626 397-7354; [irsa.ipac.caltech.edu](http://irsa.ipac.caltech.edu)

Before discussing these activities in detail, we describe first three characteristics of the archive that makes them possible:

- The breadth of data in the archive.
- The availability of services and tools, sometimes unique, that extract the maximum science content from data sets
- An archive design that is portable between platforms, extensible to many data types and wavelengths, and supports in real time queries to both very small and very large data sets.

## 2. AN OVERVIEW OF THE CHARACTERISTICS OF IRSA

### 2.1 The Breadth of Data IRSA's Holdings

When IRSA began operations in 1999, it served the data sets from IRAS, the first all-sky survey in the thermal infrared, and provided database management support to the 2 Micron All-Sky Survey (2MASS), which was then beginning observations. Since then, the archive has grown into a true multi-mission archive. These all-sky data sets remain of exceptional value in astronomy. Altogether, the archive hosts data from 24 projects and missions, 120 source catalogs, 48 image data sets, and seven spectroscopic data sets. Table 1 summarizes the major holdings and their sizes.

Table 1: Summary of IRSA's Major Holdings (May 2008)

	Wavelength Range	# of sources	# of images	# of spectra
<b>Missions/Projects</b>				
2MASS	1.2-2.2 $\mu\text{m}$	1,862,437,977	14,567,539	...
IRAS	12-100 $\mu\text{m}$	2,433,047	50,902	...
MSX	8.3-21 $\mu\text{m}$	545,274	...	...
Spitzer Legacy/First Look	0.3 $\mu\text{m}$ - 214 mm	134,229,130	14,501	17,459
IRTS	1-700 $\mu\text{m}$	14,294	1,067	2,144
SWAS	539-616 $\mu\text{m}$	...	...	63,928
<b>Value-added Data</b>				
DENIS	0.82 -2.2 $\mu\text{m}$	355,220,325	...	...
USNO-B	0.4-0.9 $\mu\text{m}$	1,045,175,762	...	...
SDSS – DR6	0.35 -0.89 $\mu\text{m}$	377,436,996	...	...
<b>Contributed Data</b>				
COSMOS	1.2 $\text{\AA}$ - 214 mm	804,451	6,555	...
IRAS: EIGA, MIGA, IRIS	12-100 $\mu\text{m}$	...	16,604	...
Spitzer/IRAC Gal. Center	3.6 -8.0 $\mu\text{m}$	1,065,565	...	...
ISO SWS	2.4-45.4 $\mu\text{m}$	...	...	17,668
<b>TOTALS</b>		<b>3,779,362,821</b>	<b>14,657,171</b>	<b>101,199</b>

Figure 1 shows a spatial coverage map of the image data sets, excluding IRAS and 2MASS. Of special interest are the Galactic Plane surveys conducted by two Spitzer Legacy projects (deemed of exceptional long-term value in astronomy), Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) and MIPS GAL, a 24  $\mu\text{m}$  and 70  $\mu\text{m}$  survey of the Galactic Plane with the Multiband Imaging Photometer for Spitzer (MIPS), and the Midcourse Space experiment (MSX) survey. Together with 2MASS and IRAS coverage in the Galactic Plane, these data sets cover the wavelength range from 1.2  $\mu\text{m}$  to 100  $\mu\text{m}$  and offer unsurpassed wavelength coverage and sensitivity for studying star formation in our Galaxy. In addition, IRSA hosts data from 9 other of the 33 total Spitzer Legacy projects (<http://ssc.spitzer.caltech.edu/legacy/>). IRSA also hosts the 2MASS Extended Mission data, which augments the 2MASS all-sky survey with multi-epoch observations and deep observations of selected areas, such as  $\rho$  Ophiuchi and the calibration fields.

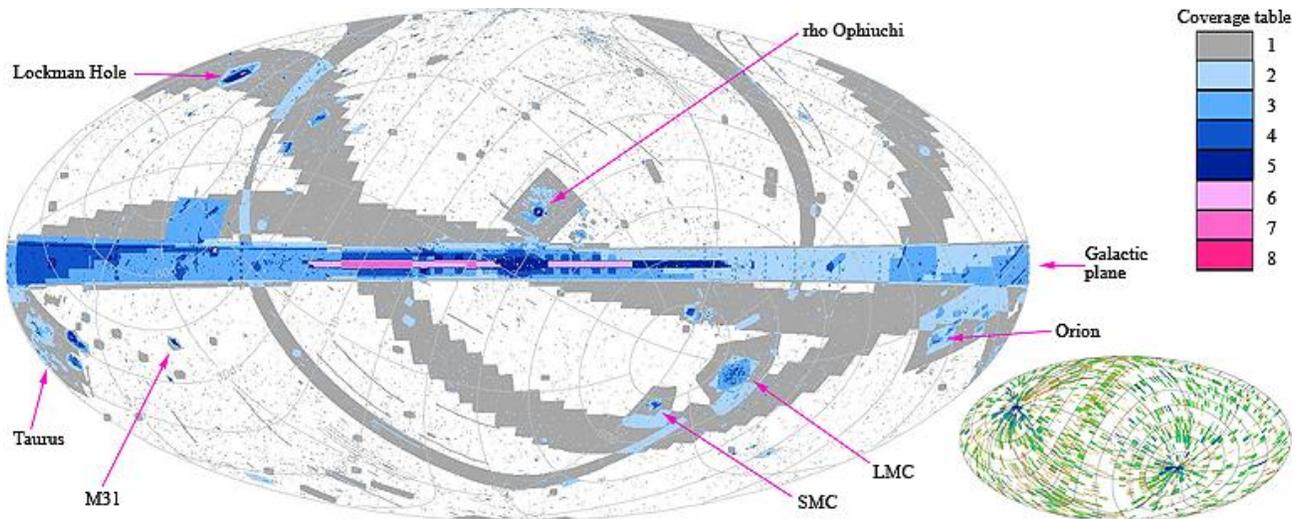


Figure 1: Sky coverage of IRSA's image sets as shown in Galactic coordinates. All-sky image sets (2MASS and IRAS, have been excluded for clarity, and images associated with the 2MASS Extended Mission multi-epoch coverage are shown in the lower right. The coverage table (upper right) shows the color mapping of the main figure, grey representing coverage by one data set and dark pink representing coverage by eight data sets. (For the inset at lower right, regions of higher coverage are shown in darker, bluer colors.) Regions of special astrophysical interest on the main figure are marked.

Third-party data sets are served when they enhance the science content of the archive and because they support the interpretation of the infrared data sets, in providing the spectral energy distributions and in discriminating astronomical sources from image artifacts. Thus, to support the interpretation of 2MASS data, IRSA has ingested the Deep Near Infrared Survey of the Southern Sky (DENIS) source catalog and United States Naval Observatory (USNO)-B catalogs. The DENIS catalog contains measurements over 16,700 square degrees of the southern sky at i, J and K. The USNO-B catalog covers the full sky at B, R, and I, and includes proper motions measured over a period of 40 years. The volume of the holdings is 23 TB, and the physical size of the archive is 82 TB, including redundant storage of data, backups, documentation, ancillary data such as indices needed for fast searches, and metadata records, needed to understand the content and provenance of the science data. Three reprocessed image data sets improve on the sensitivity and spatial resolution of the IRAS data sets: they are the Extended Infrared Galaxy Atlas (EIGA), the Mid-Infrared Galaxy Atlas (MIGA) and the Improved Reprocessing of the Infrared Sky (IRIS) data set.

## 2.2 Archive Services and Tools

*Services:* IRSA offers two types of query mechanisms for users: web interfaces and program interfaces. Both types of services provide fast access to catalogs, images and spectra, regardless of the size and complexity of the data sets, and support bulk download of large data sets. Many of these services access remote data sets to support evaluation and interpretation of infrared data. Web interfaces accept query inputs through a web browser and return results rendered on a web page. Program interfaces encode the query input in a text string that is embedded in a program and return results in a form optimized for use by a computer or by returning only data files. The program interfaces come in two flavors. One set, compliant with Virtual Observatory (VO) protocols, is intended to support large-scale discovery of data across multiple archives. Given the generic quality of these VO-interfaces, IRSA defines custom interfaces to support refined queries on all parameters of a data set. For example, a single query for Spitzer data would return all data measured with the Infrared Spectrograph (IRS) in its high-resolution modes and with the Multiband Imaging Photometer for Spitzer (MIPS) in its Spectral Energy Distribution mode. In the past two years, the use of program interfaces has accelerated faster than the use of web queries in the past to years, to the extent that in 2008 Q1, queries from program interfaces for the first time exceeded those from web queries. Fig 2 shows this effect clearly.

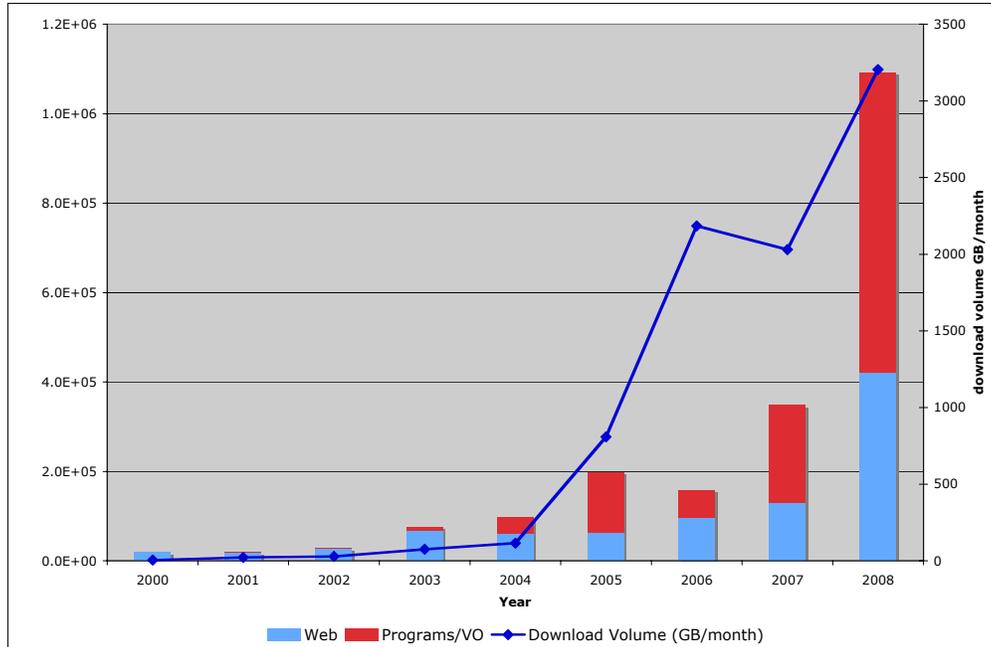


Figure 2: The growth in usage of IRSA from 2000 to 2008 (Jan – Mar) as measured by the number of queries made through web and program interfaces (including VO interfaces) and by download volume.

**Tools:** Beyond services that provide access to data holdings, IRSA hosts specialized tools that optimize the science return from these data sets. Three of them are described briefly below.

**Scanpi and HIRES:** IRSA maintains two tools, the Scan Processing and Integration tool (Scanpi) and High RESolution (HIRES), that extract higher quality data than contained in the IRAS project data sets. Completely modernized in 2007, Scanpi offers a factor of 2-5 gain in sensitivity over that of the IRAS Point Source Catalog by performing one-dimensional scan averaging of raw survey data. HIRES uses a Maximum Correlation technique to return IRAS survey images with higher resolution than the 4-arcminute IRAS Sky Survey Atlas (ISSA). HIRES default images have 2-arcminute resolution, and the algorithm allows for iteration that can achieve well below 1 arc minute, depending on wavelength and survey coverage.

**Montage:** The Montage image mosaic toolkit aggregates input FITS files into science-grade mosaics that preserve the calibration fidelity and astrometric accuracy of the input images. The mosaics meet user-specified parameters of projection, coordinates, and spatial scale. Montage supports all projections and coordinate systems in use in astronomy, and will rectify background radiation from the instrument and the atmosphere to a common level across all the input files. It has been downloaded over 1,800 times since January 2007. An on-request mosaic service, powered by Montage, generates 2MASS mosaics from the full set of survey images, the superset of the images from which the 2MASS all-sky image atlas was drawn. Additionally, it returns 2MASS, SDSS, and DSS images all sampled at the same scale. For multi-epoch 2MASS data it provides co-additions, thus providing images with the highest possible sensitivity. The multiple wavelengths covered by SDSS and DSS further support interpretation of sources and features evident in the mosaics.

### 2.3 Archive Architecture

IRSA offers considerable economies of scale and technical expertise to observatories or missions wishing to deploy archives. Projects may take advantage of a reusable Science Information System (SIS), a software architecture, a user support system, a configuration management system, data ingestion and data validation tools and methodologies, preview generation utilities, a test bed and test methodologies, built-in compliance with Virtual Observatory standards and protocols, and secure access to protected data sets.

Some remarks about the reusable architecture mentioned above are appropriate. Underpinning IRSA is a software architecture designed to support high throughput of data, ease of maintenance, and extensibility to new projects and datasets. The key design feature is modularity. All software components are written as modules or libraries that perform the tasks needed to formulate and process queries and return results. Examples are: perform coordinate transformations, prepare a query for submission to the database, filter results from a database query, and generate an HTML result page. These modules are by design general purpose and written as stand-alone modules, generally in 'C' for performance and portability. The query forms seen by users through their browsers are thin front-ends that sit atop the software architecture. While built originally to support archiving of infrared data sets, the architecture is organized to serve the types of data used by astronomers and consequently has no restriction on wavelength. New user applications take advantage of the archive infrastructure: they simply plug together existing modules, and new functionality is built as needed. Thus the architecture enables re-use of software and controls maintenance costs.

An important characteristic of the IRSA architecture is that queries on massive data sets are performed quickly: access times of a fraction of a second apply even on data sets containing a billion records. This performance is achieved through three design choices involving the layout of data, choice of database, and database hardware. All data or attributes subject to querying are organized inside the database as simple, flat tables, indexed according to a sky-partitioning scheme. Queries to these tables are made through the Informix Database Management System (DBMS), selected because of its capability to run parallel queries on the fly. The DBMS resides on dedicated servers and connected to high-speed database mass storage disks through multiple fiber-channel connections.

### 3. APPLICATIONS OF IRSA TO OBSERVATORIES AND MISSIONS

#### 3.1 Deploying archives for observatories and missions

This section describes archives built with the IRSA archive infrastructure on behalf of observatories and missions.

*2MASS and The Wide-field Infrared Survey Explorer (WISE):* IRSA developed its architecture to support database management, product generation and data dissemination for the 2MASS mission. The 2MASS archive was in fact built in situ inside IRSA, and public release simply involved removing protections on data access services. WISE (<http://wise.ssl.berkeley.edu/>) is scheduled for launch in 2009 and will survey the sky at 3.3  $\mu\text{m}$ , 4.7  $\mu\text{m}$ , 12  $\mu\text{m}$  and 23  $\mu\text{m}$ . It will use the same archive model as 2MASS. The archive subsystem at the WSDC stores raw and processed survey data that is intended to enable distribution to the WISE project and the astronomical community. Raw telemetry is stored on local magnetic disk for the duration of the mission, and written to magnetic tape for long-term archiving. The processed image and extracted source data will be integrated into IRSA. Access to the processed data products by the WISE team during mission operations, and by the public to the WISE Catalog and Image Atlas is via web and program services, including VO compliant services, which will be developed and maintained at IRSA. These include image preview, retrieval and inventory services, powerful catalog and database query engines, as well as services that provide interoperability with other IRSA catalog and image holdings. IRSA will assume responsibility for the maintenance of the archive at mission end in 2012, under current schedules.

*The W.M. Keck Observatory Archive (KOA):* The project is funded by the National Aeronautics and Space Administration (NASA) as a collaboration between the Michelson Science Center (MSC) and the W.M. Keck Observatory. Currently, it archives data from the High Resolution Echelle Spectrograph (HIRES) [1]. This instrument supports active research in planetary astronomy and quasars, but is perhaps most celebrated for its role in the discovery of extra-solar planets. The goals of the archive are: promote NASA's science theme of searching for extra-solar planets; curate and disseminate observations made on Keck Single Aperture instruments to maximize scientific return from the observatory, and enable long-term instrument performance studies that will benefit the development of observing programs.

The KOA (<http://msc.caltech.edu/archives/koa/>) opened for business in July 2006, serving unprocessed science and calibration data ("level 0") from the upgrade to HIRES deployed in August 2004; this upgrade replaced the original CCD chip with a three CCD mosaic that offered superior throughput in the blue and violet. As of May 2008, the KOA serves 693 nights of data measured with the upgraded instrument. These data are 2.4 TB in size, comprising 65,000 calibration files and 35,000 science files, each containing measurements for all three CCD's. Principal investigators have proprietary access to their data sets *on a per CCD basis* for at least 18 months after the date of observation.

Currently, 105,000 science and calibration individual CCD files are public. Since opening to the public in July 2006, the KOA has received 48,00 queries and 1.3 TB of data have been downloaded from it.

The KOA has recently ingested ten years worth of observations over 1261 nights made with the single-chip CCD. These data comprise 47,193 science files and 47,209 calibration files, a total volume of 380 GB. Observations made between 2001 and 2004 are already public, and the rest will become public on June 30, 2008. Finally, in Summer 2008, the KOA will release an extracted (“level 1”) browse product that will provide a guide to the science content of each data set.

The KOA uses a modular design consisting of four uncoupled components, as follows:

- Data Evaluation and Preparation (DEP), performed at the Observatory)
- Trans-Pacific Data Transfer (TPX) from the Observatory to MSC
- Science Information System (SIS), maintained at IPAC
- User Interface, maintained at MSC.

Figure 3 shows the relationships between the components.

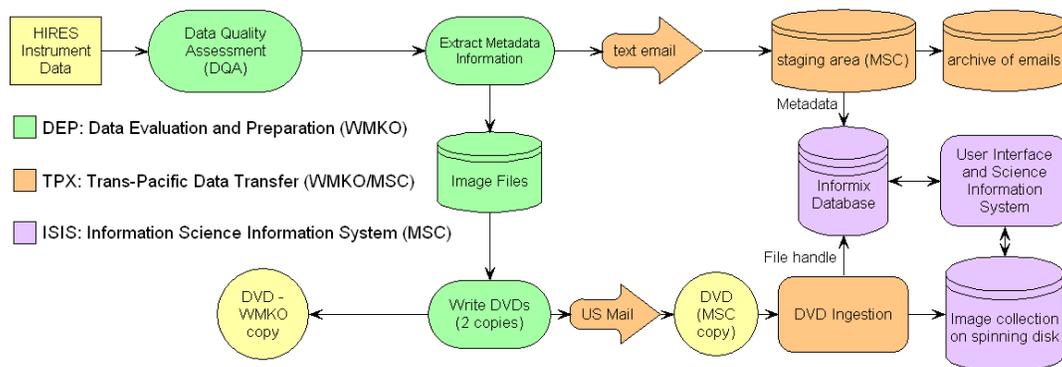


Figure 3: The Components of the Keck Observatory Archive (KOA)

This modular design overcomes the geographical separation between the telescope and the archive, reduces maintenance costs (as each component can be upgraded independently of the others), and enables the KOA to inherit the IRSA architecture. The functions of each component are as follows. DEP evaluates the FITS file headers for compliance with FITS standards and for completeness and legal ranges, and then converts them to a metadata table. TPX then sends the metadata to MSC via daily e-mail and ingested into the archive. The FITS files are written to DVD and shipped weekly to MSC. The metadata and FITS files are validated by comparing their checksums calculated at MSC with those delivered with the data. The SIS, inherited from IRSA, provides the functionality needed to support the archive, and the user interface accepts queries and returns results to the user.

*The NASA Star and Exoplanet Database (NStED):* NStED (<http://nsted.ipac.caltech.edu/>) is an archive of positions, kinematics, photometry, stellar parameters, exoplanet parameters, images, spectra, photometric light curves, and radial velocity curves for stars known to host exoplanets and for stars that are candidates for exoplanetary studies. In particular, NStED maintains an archive of ground-based photometric light curves from surveys intended for the discovery of exoplanetary transits. In recent years, the ground-based surveys have greatly improved their photometric and detrending routines to enable the discovery of transiting exoplanets. These data sets provide a valuable resource for stellar variability and asteroseismology studies. Most of these data remain unavailable to the astrophysical community, and standardization between data sets remains elusive. The extant ground surveys provide unique data sets for the identification and study of variable sources in these large surveys. Five ground-based survey programs, shown in Table 2, have been ingested and provide public access to the survey light curves and any associated data. On average, the

surveys contain 20,000–50,000 stars with hundreds to thousands of photometric epochs per star. NStED will investigate, in collaboration with these groups, archiving images to help validate rare events, such as transits. New transit data sets are scheduled for ingestion into the archive in the next three years, per agreements with survey programs.

Table 2: Transit Data Sets Served Through NStED

Survey Region	Number of Stars	Time-Series Filter	Time Span (days)	Number of Epochs
TrES-Lyr1	25,947	r, R	75	~15,000
NGC 2301	3,961	R	14	~150
NGC 3201	58,666	V, I	700	~120
M 10	43,930	V, I	500	~50
M 12	32,378	V, I	500	~50

As with KOA, NStED is built on the IRSA architecture. For example, it shares with IRSA tools for accessing the database, tools for web based plotting, task management tools (to prevent browser timeouts), and an object name look-up service that was built as an extension of IRSA’s tool, so all nomenclature is resolved through one engine. The architecture was extended to enable the user interface to built dynamically from configuration tables; new parameters can then be added without coding. Figure 4 shows the transit of the exosolar planet TrEs-1 in the light curve visualizer, an x-y plotter inherited from IRSA.

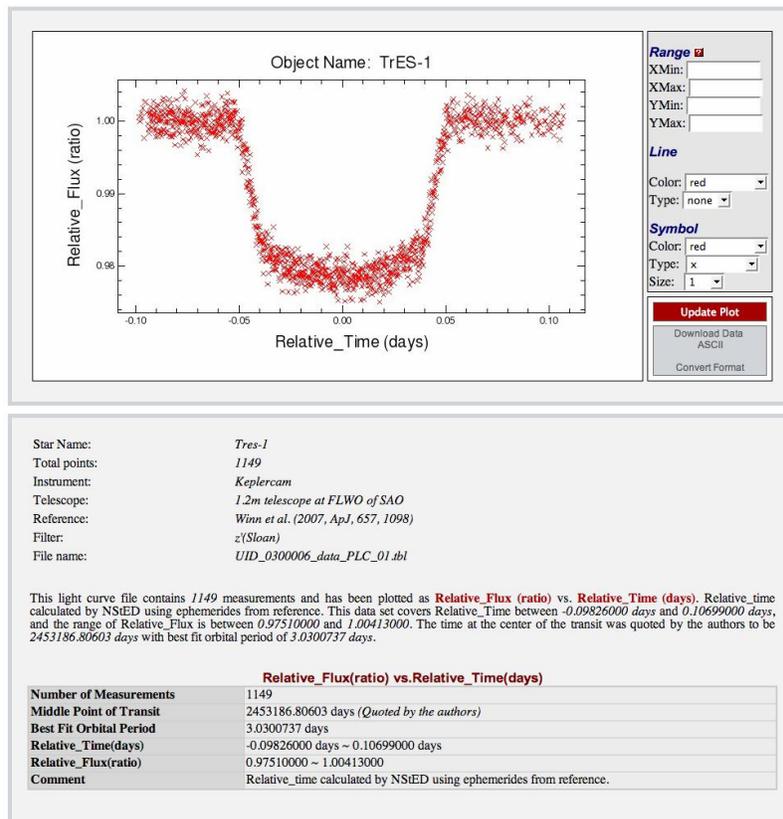


Figure 4: The light curve of the exosolar planet TrES-1 as it transits its host star, as shown in the light curve viewer inherited by NStED from IRSA. It is a web-based interactive viewer that shows the attributes of the data set under visualization.

### 3.2 Observation planning

The Spitzer and Herschel observation planning tools, *Spot* and *HSpot*, retrieve through IRSA's program interfaces images from 2MASS, IRAS, and MSX and IRIS, and sources from the 2MASS catalogs to aid in targeting sources, avoiding bright interfering sources, and finding guide stars, and to predict the infrared sky brightness as observed by Spitzer. Spitzer observers additionally use Scanpi in observation planning: it provides the most sensitive measurements that can be obtained from the IRAS data sets.

The Michelson Science Center (MSC) has created a tool, *getCal*, to support interferometric observation planning. It is the standard tool for planning and observation scheduling for the Palomar Testbed Interferometer and the Keck Interferometer; it is also used at CHARA and VLTI. This tool accesses 2MASS and IRAS source catalogs through program interfaces to select guide stars and targets for interferometric measurement. The 2MASS fluxes are used to model the spectral energy distributions for physical analysis of targets. The 2MASS and IRAS data are crucial elements in reliable identification of stars with circumstellar dust shells or strong emission components since optical measurements cannot be reliably extrapolated to longer wavelengths where the interferometers operate. Without such vetting, roughly half the calibrator observations would be unusable.

### 3.3 Support pipeline processing

The *Spitzer Space Telescope* exploits the high astrometric accuracy of sources in the 2MASS point source catalog to improve the pointing accuracy of sources in mid-infrared images. This is requisite to obtaining the best possible signal-to-noise and image co-registration in the Spitzer science products [2]. The Spitzer pipeline accesses the 2MASS point source catalog on-the-fly to obtain the requisite astrometric information for 2MASS images.

### 3.4 Support generation of new products.

Three Spitzer Space Telescope Legacy projects (deemed of exceptional long term value to astronomy) have incorporated Montage into their image product generation pipelines, and another is using it to investigate the optimum image plate scale needed to cross-identify extended protostellar cores and infrared dark clouds in multi-wavelength mosaics of the Galactic plane. The HIRES tool has been used to generate IRAS products that have higher spatial resolution than the nominal mission products. The Extended-Infrared Galaxy Atlas (EIGA) is an extension of the original IRAS Galaxy Atlas (IGA) at 60  $\mu\text{m}$  and 100  $\mu\text{m}$  to more northerly galactic latitudes. The Mid-Infrared Galaxy Atlas (MIGA) is a high-resolution image atlas of the Galactic plane at 12  $\mu\text{m}$  and 25  $\mu\text{m}$ , and is an adjunct to the Infrared Galaxy Atlas.

### 3.5 Support for New Missions and Projects

In addition to building its archive in situ within IRSA, WISE has been using Montage to support its mission preparations:

- To provide an independent check of the specialized coadding/mosaicking software that has been developed for the WISE pipelines, and
- To produce large-scale image mosaics of 2MASS and Spitzer data of the ecliptic polar regions. These regions are the WISE "touchstone" fields that will be observed on nearly every WISE orbit, and contain the primary WISE photometric calibration star network. 2MASS and Spitzer mosaics are being used to characterize the regions and to assess usability of the WISE standard stars from the standpoint of environment, crowding, and proximity to other bright objects that might complicate WISE measurements.

In operations, WISE will use Montage in two ways:

- To construct smaller mosaics of 2MASS image data for the purpose of WISE processing validation and science QA review. For example, to compare WISE image data with 2MASS images in the same regions to assess artifact identification and moving object association and associated inertial source confusion.
- To underpin the WISE/IRSA image access service to allow users to request Atlas Images on any specified center, project, pixel scale and size. The WISE Atlas Images will be constructed on a predefined grid, with a pixel scale approximately two times finer than the raw WISE detector pixel scale (4k x 4k pixels at a sampling of 1.375"/pix, covering 1.56 x 1.56 deg). The WISE/IRSA image server will provide cutouts (sub-images) or

mosaics if the requested footprint covers more than one WISE Atlas Image, and will support user specified projections and spatial samplings.

### **3.6 Proposal Calls**

IRSA has built web-based submission systems for the bi-annual NASA-Keck observing time call, which has approximately 45 responses per semester, and for the SIM/PQ Planet-Finding Astrometry Analysis Teams call. The web tools take advantage of validation of the ranges and formats of input data built into IRSA. The submissions are maintained in a database, where proposal evaluators can access them using IRSA's tools.

### **ACKNOWLEDGEMENTS**

I wish to thank Dr. Rachel Akeson, Dr. Roc Cutri, Dr. Dawn Gelino, and Dr. Mark Lacy for discussions, and the members of the IRSA team for their dedicated work: A. Alexov, N. -M. Chiu, J. Good, J. D. Kirkpatrick, M. Kong, A. Laity, D. McElroy, S. Monkewitz, A. Zhang. IRSA is supported by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

### **REFERENCES**

- [1] Berriman, G. B., Ciardi, D. R., Laity, A. C., Tahir-Kheli, N., Conrad, A., Mader, J., and Tran, H. "The Design of the W. M. Keck Observatory Archive," Shopbell, P. L., Britton, M. C., and Ebert, R. [Astronomical Data Analysis and Software Systems XIV], ASP Conference Series, Vol 347, 627-631 (2005)
- [2] Laher, R., McCallon, H., Masci, F., and Fowler, J. "Position Refinement of Spitzer Space Telescope Images," Gabriel, C., Arviset, C., Ponz, S., and Solano, E. [Astronomical Data Analysis and Software Systems XV], ASP Conference Series, Vol 351, 169-172 (2006)