# Data Quality Analysis at the Spitzer Science Center

Vincent Mannings*, Russ R. Laher

Spitzer Science Center, California Institute of Technology, Pasadena, CA 91125

## ABSTRACT

Data Quality Analysis (DQA) for astronomical infrared maps and spectra acquired by NASA's Spitzer Space Telescope is one of the important functions performed in routine science operations at the Spitzer Science Center of the California Institute of Technology. A DQA software system has been implemented to display, analyze and grade Spitzer science data. This supports the project requirement that the science data be verified after calibration and before archiving and subsequent release to the astronomical community. The software has an interface for browsing the mission data and for visualizing images and spectra. It accesses supporting data in the operations database and updates the database with DQA grading information. The system has worked very well since the beginning of the Spitzer observatory's routine phase of operations, and can be regarded as a model for DQA operations in future space science missions.

Keywords: data processing, data quality analysis, data archive, data mining, Spitzer Space Telescope

## 1. INTRODUCTION

The Spitzer Space Telescope is the fourth and final NASA Great Observatory. It has been operating successfully since it was launched in August 2003 and includes three instruments for sensitive astronomical infrared observations: the Infrared Array Camera (IRAC), the Multiband Imaging Photometer for Spitzer (MIPS), and the Infrared Spectrograph (IRS).[1] Only one instrument is operated at any given time, and observations are conducted in campaigns lasting between one and three weeks. The baseline instrument campaign schedule comprises a week or two of IRAC measurements, followed by a MIPS campaign of similar length, and then an IRS campaign before returning once again to IRAC.

A set of astronomical data obtained with IRAC, MIPS or IRS for a given user's science program is typically comprised of a contiguous sequence of measurements requiring several tens of minutes of Spitzer-Space-Telescope observing time ($\mu$=30.3 minutes, $\sigma$=45.8 minutes), and normally a small subset of the data taken during the associated instrument observing campaign. This observing sequence is formally referred to as an Astronomical Observation Request (AOR) in Spitzer parlance. An AOR is composed of multiple Data Collection Events (DCEs), where a DCE constitutes one or more raw science images obtained with a Spitzer instrument (depending on the instrument and observation mode).

The raw images or DCEs collected during the observations, along with pointing and housekeeping data are transmitted daily via telemetry to earth from the Spitzer observatory. The raw data undergo level-0 processing at the Jet Propulsion Laboratory (JPL) to produce raw FITS images or FITS image data cubes[2], one per DCE, and are then sent to the Spitzer Science Center (SSC), the institutional home for processing, calibrating, archiving, and distributing Spitzer raw data and processed products. The headers of the FITS images contain essential information about the DCEs. Pointing-history files that cover 12-hour time periods are also received at the SSC from JPL; these files are used later to assign celestial sky positions to all DCEs. The housekeeping data are time-sampled positions, temperatures, voltages, etc. that can be trended to further analyze information about the science instruments and their observations.

* vgm@ipac.caltech.edu

A high-level Spitzer project requirement calls for quality analysis of the calibrated astronomical data. This is an important part of the processing that occurs at the SSC. The data quality analysis is done from a science perspective and answers the following paramount question: do the data for a given AOR conform reasonably well to the expected performance of the Spitzer Space Telescope for the measurements that were scheduled by the observer? Ultimately, whether a given AOR can also be employed by the user to address the science goals of the observer's program is a question that must be answered by the user. The quality analysis is performed at the SSC by a team of four specialists in one work shift (nominally 40 hours per week per specialist), and it is carried out between the time the data are initially processed and before the data products are finally copied to the Spitzer public archive for retrieval by the end user. We briefly describe in this paper the automated pipeline data processing, the subsequent data quality analysis operations, and the software tools and infrastructure that support DQA operations.

## 2. AUTOMATED PIPELINE DATA PROCESSING

Figure 1 gives the flow of Spitzer data through the SSC. The DCE FITS images, pointing-history files and housekeeping-data files are ingested at the SSC as they are received from JPL. This involves three basic steps: 1) registering the files in the Spitzer operations database, 2) storing the files in the Spitzer archive, and 3) applying read-only file permission. The DCEs are then automatically processed and calibrated using a suite of software pipelines that were developed and are periodically upgraded by the SSC downlink software group. Since the housekeeping data are not needed for the routine processing, it is ingested into a separate database.

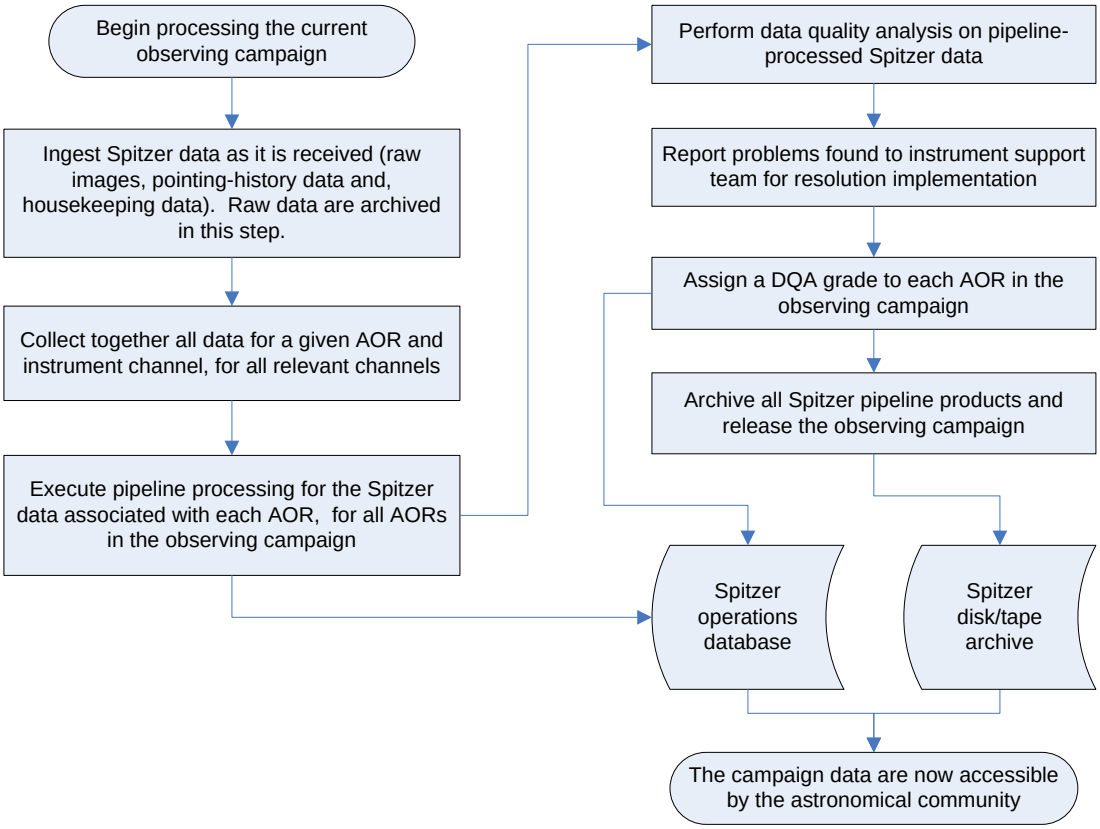


Fig. 1. Data processing and DQA activities at the Spitzer Science Center.

The data calibration actually has two phases. New data are initially processed using "fallback" calibration data[3]. This occurs shortly after ingesting and collecting together the data to form a complete AOR. Once all AORs in the entire instrument campaign have been initially processed, the campaign's full set of calibration data are then used to fine-tune the calibration of all the data acquired in that campaign. This is accomplished by a second round of processing, typically within a week of the end of an observing campaign.

The pipeline processing is done on copies of the DCEs from the Spitzer archive (the original DCEs received from JPL are kept pristine in the Spitzer archive). The SSC downlink software system generates via pipeline processing for each DCE a Basic-Calibrated-Data (BCD) product, which includes image data that have been converted to absolute flux-density units (MJy/sr or Mega-Janskys per steradian) by the flux-calibration process, and WCS-projection parameters[4] written to their FITS headers as derived from the pointing-history data. As examples, Brandenburg *et al.* (2005) describe the SSC's IRAC science and calibration pipelines[5] and Masci *et al.* (2005) describe the SSC's MIPS 24-μm image data processing[6].

After BCD processing, ensembles of BCDs[7] are further pipeline processed *en masse* to create so called "post-BCD" data products. Typically an ensemble is comprised of all DCEs of the same instrument channel in an AOR, where an instrument channel corresponds to image data limited to within a specific infrared spectral passband. However, in some cases, ensembles span multiple instrument channels, such as when computing a common WCS reference frame for re-sampling BCD images, which is necessary for multi-color combining. Examples of post-BCD data products are as follows: re-sampled, co-added images to reduce image-data noise; mosaics of re-sampled, co-added images for mapping areas of the celestial sky that are much larger than an instrument's field of view; catalogs of point sources derived for separate instrument channels; and "band-merged" lists of point sources detected in common across multiple instrument channels. These advanced data products, as generated directly by SSC automated pipelines, are of sufficient quality to be published directly in peer-reviewed astronomy and astrophysics journals.

Throughout the ingest stage, the BCD-pipeline-processing stage, and the post-BCD-pipeline processing stage, several quantities that are useful in the data quality analysis process are computed and/or stored in various tables in the Spitzer operations database. For example, information about missing lines of image data and missing ancillary-data packets (ancillary data are packaged in the DCE's FITS header) is stored in the dceInstances database table when the DCE is ingested. At the beginning of the BCD pipeline, various image statistics for the DCE are computed and associated with the best instance of the DCE via the database index "dceId" when stored in the database. At the end of the BCD pipeline, similar image statistics for the associated BCD are computed and associated with the data-product index "dpId" when stored in the database. Finally, at the end of the post-BCD pipeline, image statistics, outlier counts, and other DQA measures are computed for the post-BCD products and associated with the ensemble-product database index "epId" when stored in the database. A general image software program called QATOOL, which was developed by one of the authors (R. R. L.), is used in the pipelines to routinely compute image statistics.

## 3. QUALITY ANALYSIS OF SPITZER DATA

Following automated pipeline data processing, the key DQA activities for Spitzer raw and processed science data are listed as follows (see the right-hand side of Figure 1):

1. A team of data quality analysts sample the data from each AOR to search for anomalies and grade the data;
2. The data are reprocessed at the end of the campaign, and permanently stored in a publicly accessible data archive for retrieval by users worldwide; and
3. Upon retrieval by a user, the SSC's archive retrieval tool (Leopard) triggers the construction of a text-formatted README file that includes the results of the DQA analysts' study of the data, which is packaged with the data downloaded by the user.

The DQA team is led by one of the authors (V. M.), who is a Ph.D. astrophysicist with 20 years of astrophysics research and operations experience. The other group members include a Ph.D. space physicist with a decade of experience in the analysis of data from space-based Earth climate observatories, and two staff members with M.S. degrees in astronomy.

During regular business hours, Monday through Friday, the DQA team examines, trouble-shoots and grades all newly downlinked and calibrated science AORs. New AORs are automatically assigned to team members (henceforth DQA analysts) by Perl scripts that query the Spitzer operations database twice each day for new products generated by the processing software. Each DQA analyst works on about 10 new AORs per day.

The analysis of each AOR includes both rudimentary checks and detailed study of any unexpected features. The basic checks performed by the DQA analyst include:

- Did the observatory collect the expected number of DCEs? The answer is affirmative for the vast majority of AORs. Of those few AORs with missing DCEs, only one or two DCEs are absent out of several hundred or several thousand expected. In the extremely rare case when more than a quarter of an AOR's DCEs are absent, an SSC procedure exists whereby a request is made for the AOR to be scheduled again for a repeat observation. This has occurred for of order 10 AORs from the approximately 19,000 science AORs acquired to date by Spitzer.

- Does a given DCE have any missing image or ancillary data? DCEs are version-controlled in the Spitzer operations database. If the first version has any missing data, JPL builds another version automatically after the DCEs with missing data are re-transmitted from the spacecraft, which almost always occurs within 12 or 24 hours. For 99.21% of the DCEs, there are no missing data in the first version. Throughout the mission so far, only 0.0186% of the final versions of the DCEs actually have permanently missing data, which indicates that Spitzer's onboard data storage and telemetry systems are performing superbly.

- Did the SSC processing system generate the expected number of BCDs and post-BCD products? The answer is affirmative for more than 99.99% of AORs. A request for reprocessing to the SSC-processing operations team typically yields any missing and desired products within a few days.

The DQA analyst then proceeds to examine and report on the image data comprising an AOR. The tools used for display and analysis are described in the next section. Here, we focus on the analysis process itself. This process has three steps:

1. The data are sampled, displayed and analyzed.
2. A release status is selected and entered into the Spitzer operations database for the AOR. The statuses are described in detail later in this section. They basically take the following forms: nominal, non-nominal, or failed. Over 98% of the observed AORs to date have been declared nominal; the rest could not be declared nominal mainly because of a few short episodes of enhanced solar activity (see below for more details).
3. If necessary, the DQA analyst will record standard and/or free-form comments in the database for the AOR. Standard comments are selected from a menu, while free-form comments are employed to warn the user of any unusual features in the data.

The image-mosaic map from the post-BCD pipeline is the main product employed for DQA purposes, although the DQA software system also provides immediate access to the associated BCDs and DCEs. One map can represent of the order of hundreds of DCEs/BCDs in the case of an IRAC map, and thousands of DCEs/BCDs for a MIPS scan map. The DQA analyst retrieves and displays a map so that it can be checked for unexpected features. The features found in image data can include both the expected and unexpected varieties. For example, known and occasionally expected features in IRAC images include multiplexer-bleeding and column pull-down (both the result of observing very bright point sources), stray light from the optics, and optical banding. Such artifacts are very well documented (e.g., see http://ssc.spitzer.caltech.edu/archanaly) and are usually not noteworthy (of little or no concern); the DQA analyst will see them, but ignore them. On the other hand, features that do prompt action include, for example, elevated rates of high-energy particle hits during periods of enhanced solar activity. This has occurred during three such periods since launch, each lasting several days. High particle rates can bring down the effective sensitivity of the observations (since data with radiation hits are excluded during data reduction), occasionally necessitating a subsequent repeat observation. Again, this has fortunately happened only rarely during the Spitzer mission so far.

Another occasional, but potentially serious feature in the image data is a latent image from a very bright source observed earlier in the campaign. Latents are registered as ghost-like images in some or many BCDs and incorporated into the image-mosaic map. Note that this feature is a temporary phenomenon and is harmless to Spitzer's imaging instruments. The effects of latents are wiped from the instrument focal-plane array(s) during periodic annealing. When their signatures are found in the data, the DQA analyst will discuss the latents with the instrument's support team at the SSC, along with members of the SSC's Science User Support group, in order to determine the impact on the AOR's science goals. The impact is typically negligible. For example, the latents might be in an instrument channel that is not important to the science goals, or they might not be superimposed on the primary targets in the map. Latents occurred during most campaigns in early nominal Spitzer operations. Now, known bright sources that are likely to cause such problems are scheduled for observation at the end of the campaign.

The DQA analysts typically complete the assigning of final release statuses to all AORs in a campaign before the second round of improved calibration reprocessing is completed. Analysis and grading of AORs can, if necessary, continue during reprocessing and even after the campaign data products have been copied to the publicly-accessible data archive. The Spitzer operations and replicated archive databases always display the latest information recorded for an AOR by the DQA analyst. There are five possible final release statuses for an AOR and only one of these is assigned to a releaseable AOR:

1. NOMINAL/RELEASABLE. The data for this AOR meet the specifications of the observing program. No significant problems are identified.

2. NON-NOMINAL/NO-REPEAT/RELEASABLE. This AOR has documented problems, but the data are nevertheless useable by archive users. The AOR will not be repeated because:

    a. The data meet the expectations of the program for this particular AOR;
    b. The data do not meet the programs expectations for this AOR, but the AOR's problems have an insignificant impact on the overall program; or
    c. The AOR cannot, in any case, be rescheduled because Spitzer can no longer support the AOR for technical reasons (note that this option has not been applicable to date.).

3. NON-NOMINAL/REPEAT/RELEASEABLE. The AOR has documented problems. Its data do not meet the expectations of the program. However, it has potential utility to the archive user and will therefore be released. The AOR's problems will have a significant impact on the expectations of the program. The SSC Director's office approves resubmission of this AOR to the scheduling pool.

4. FAILED/NO-REPEAT/RELEASEABLE. The AOR's data are not useable by the archive user, and this AOR will not be repeated because either the failure of this AOR has an insignificant impact on the overall program or the AOR cannot, in any case, be rescheduled because Spitzer can no longer support the AOR for technical reasons.

5. FAILED/REPEAT/RELEASEABLE. The AOR's data are not useable by the archive user and the failure of this AOR will have a significant impact on the expectations of the program. The SSC Director's office approves resubmission of this AOR to the scheduling pool.

As of April 2006, the overall distribution of statuses for all science AORs obtained during routine Spitzer operations using all three instruments is: 98.4% nominal (19,004 AORs), 1.4% non-nominal (278 AORs) and 0.2% failed (36 AORs). Some 75 AORs have been repeated. The distribution of statuses for AORs from each instrument is very similar to the overall distribution.

As indicated earlier, a text-formatted README file is generated when a user retrieves the data for an AOR. This file displays query data retrieved from the Spitzer operations database in a "QA Summary Page". Figure 2 gives an example README file for a real IRAC science AOR observed in February 2005.

```
QA Summary Page              Date: 2006-04-19 16:16:50.000

This file provides:
1. Basic information on the request (instrument, target name, etc.);
2. The results of data quality analysis at the Spitzer Science Center (SSC);
3. An accounting of the request's raw science data (DCEs); and
4. Pipeline software-version information.

(1) DESCRIPTION OF OBSERVING REQUEST (AOR or IER)
Telescope:             Spitzer
CampaignId:            772 (IRAC006000)
ReqTypeName:           AOR
ReqModeName:           IracMap
ReqKey:                10738432

TargetName:            IRAS 05413-0104
Program Title:         H2_OUTFLOWS
Program ID:            3315
RequestTitle:          hh212_irac_map
Observation Start:     2005-02-22 16:30:56.202
Observation End:       2005-02-22 16:47:06.207
This request was a "cold" observation.

(2) DATA-QUALITY-ANALYSIS RESULTS
AOR quality status:    Nom Rlse
Please see http://ssc.spitzer.caltech.edu/archanaly/
for details of the Spitzer science archive, pipeline upgrades,
analysis tools, and instrument-specific data handbooks.

(3) SUMMARY OF REQUEST'S SCIENCE-DATA CONTENT
Expected number of DCEs:  192
Received number of DCEs:  192
Number of missing DCEs:   0
Number of received DCEs with missing FITS-header lines: 0
Number of received DCEs with missing lines in the image data: 0

(4) PIPELINE SOFTWARE-VERSION INFORMATION
IRAC Channel-1 Software Version:    S13.2.0
IRAC Channel-2 Software Version:    S13.2.0
IRAC Channel-3 Software Version:    S13.2.0
IRAC Channel-4 Software Version:    S13.2.0
```

Fig. 2. Sample README file containing useful DQA information on the data for AOR #10738432.

## 4. SOFTWARE TOOLS FOR DATA QUALITY ANALYSIS

The necessity of people as an integral part of the DQA process was recognized early in the development of the Spitzer mission and this was incorporated into the design philosophy of the DQA software system. The first rule of data, before any kind of data analysis is applied, is to “look at the data” and humans do this best. Simply put, the DQA process cannot be machine-automated and requires participation by humans who have built up a knowledge base from their experiences in looking at the data. The aforementioned SSC DQA analysts fill this role.

The DQA software system consists of a collection of command-line Perl scripts that are executed in the DQA Unix environment and web-based CGI scripts, Java applets, and dynamically-generated HTML documents that are executed and viewed in a web browser. The software was developed by the SSC Downlink team, and it was tested, deployed, and fully functioning in the SSC science-operations system before the Spitzer Space Telescope was launched in August 2003. Software upgrades have been made roughly twice each year since launch. The loose-integration architecture and scripting nature of the software allows software upgrades to be made very quickly, even by developers new to the software. Also, scripts are, by far, the easiest kind of software to patch into Spitzer operations when bugs are found. CGI.pm is a Perl package that is often used at the SSC for rapid development of CGI scripts. See Figure 3 for an example CGI script from the DQA software suite. Perl packages for accessing the Informix database are also used by the software. Where possible, the database queries are executed via pre-compiled Informix database stored functions for maximum speed.

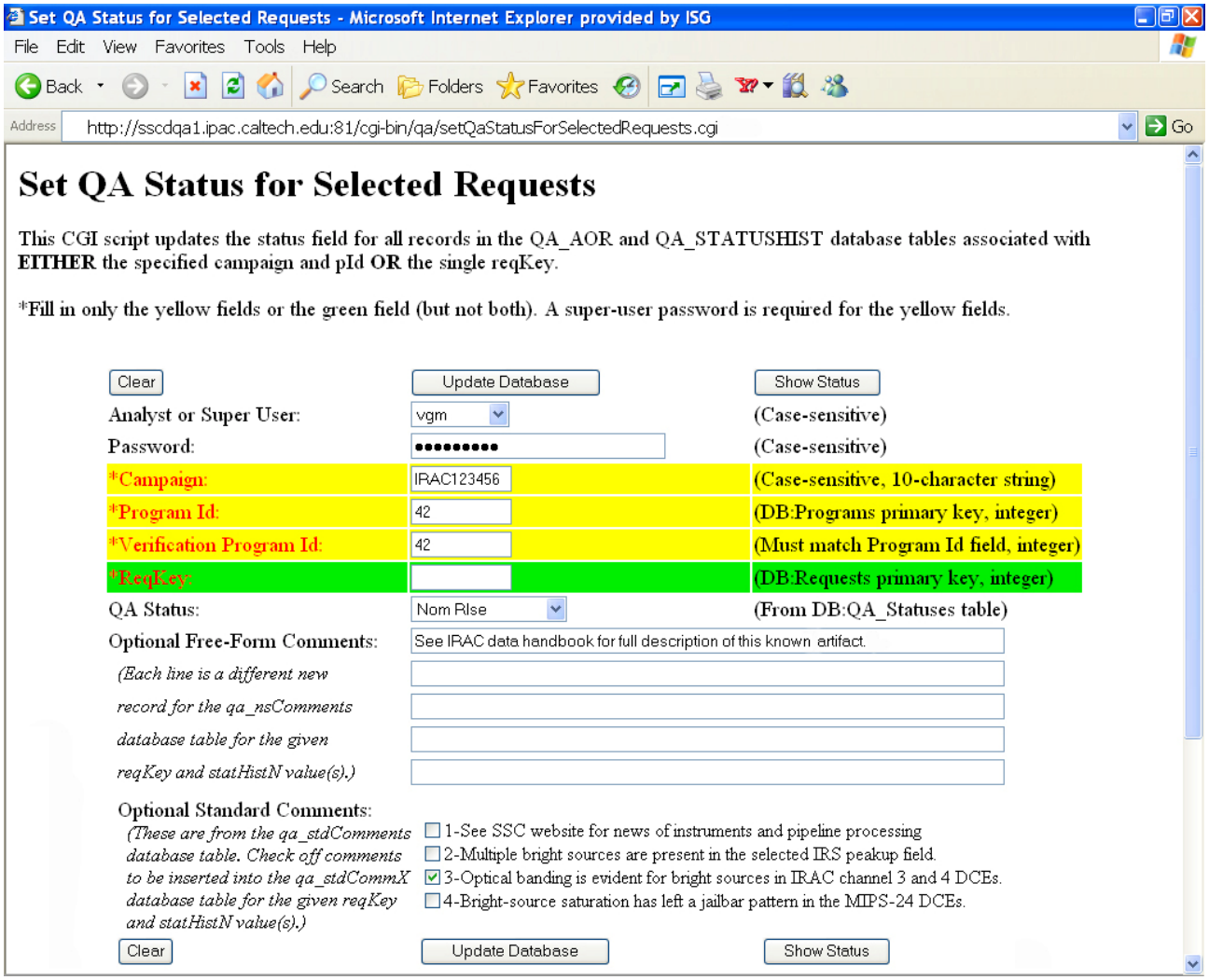


Fig. 3. Screen shot of one of the CGI scripts of the DQA software system at the Spitzer Science Center.

The Science Data Analysis Tool, or SDAT, is the primary tool used by the DQA team. SDAT can access all the DCEs, BCDs and post-BCD products comprising any Spitzer AOR. It can display images and spectra, and supports basic image analysis and histogram display. It also allows the DQA analyst to record AOR status flags and comments in the Spitzer operations database. Figures 4 and 5 illustrate some of SDAT's visualization capabilities, using as examples a channel-1 IRAC map of the astronomical source IRAS 05413-0104 and its channel-3 counterpart.

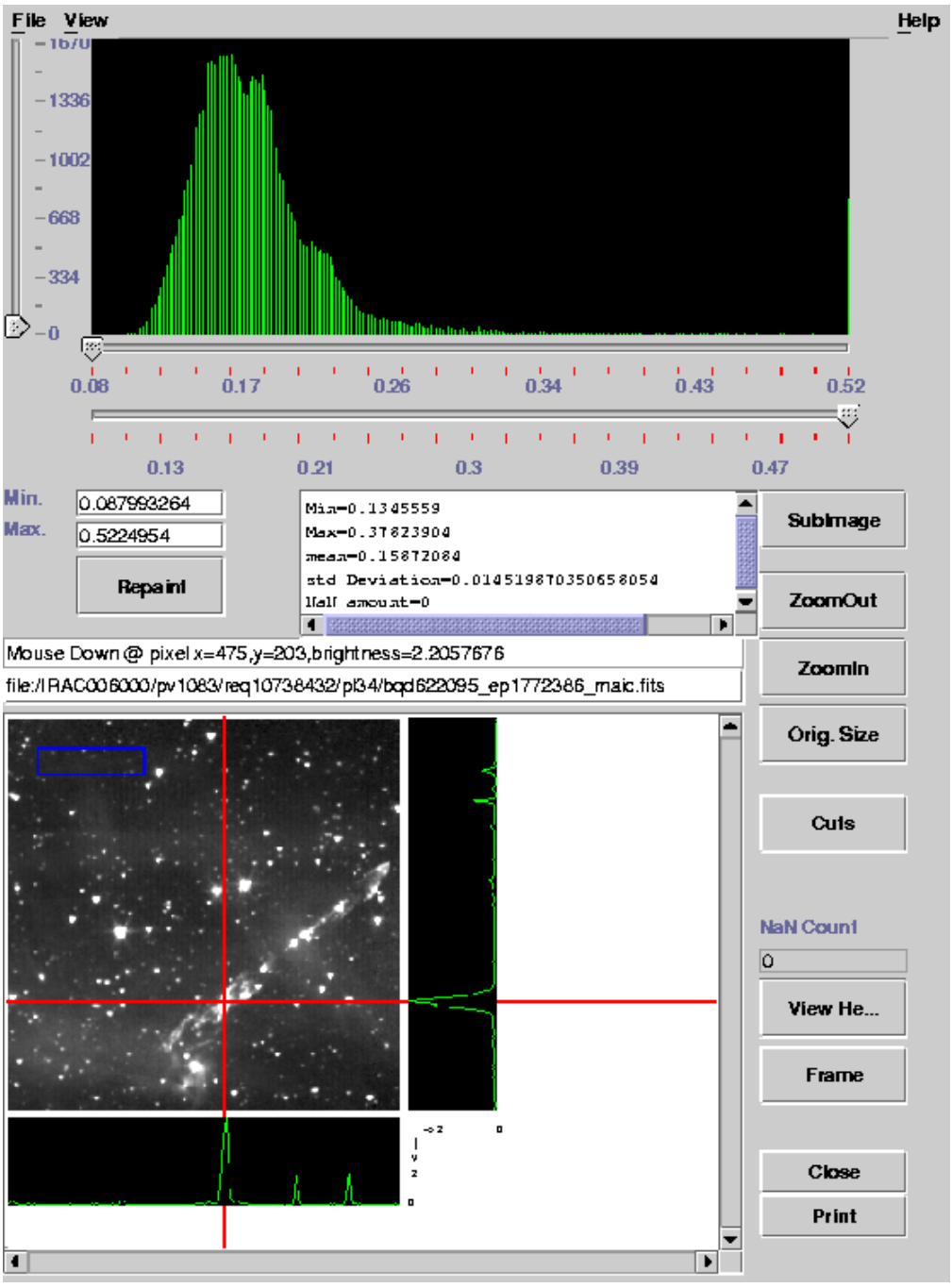


Fig. 4. Screen shot of SDAT, a component of the DQA software system at the Spitzer Science Center. Shown is a channel-1 IRAC map. See Figure 5 for its channel-3 counterpart.

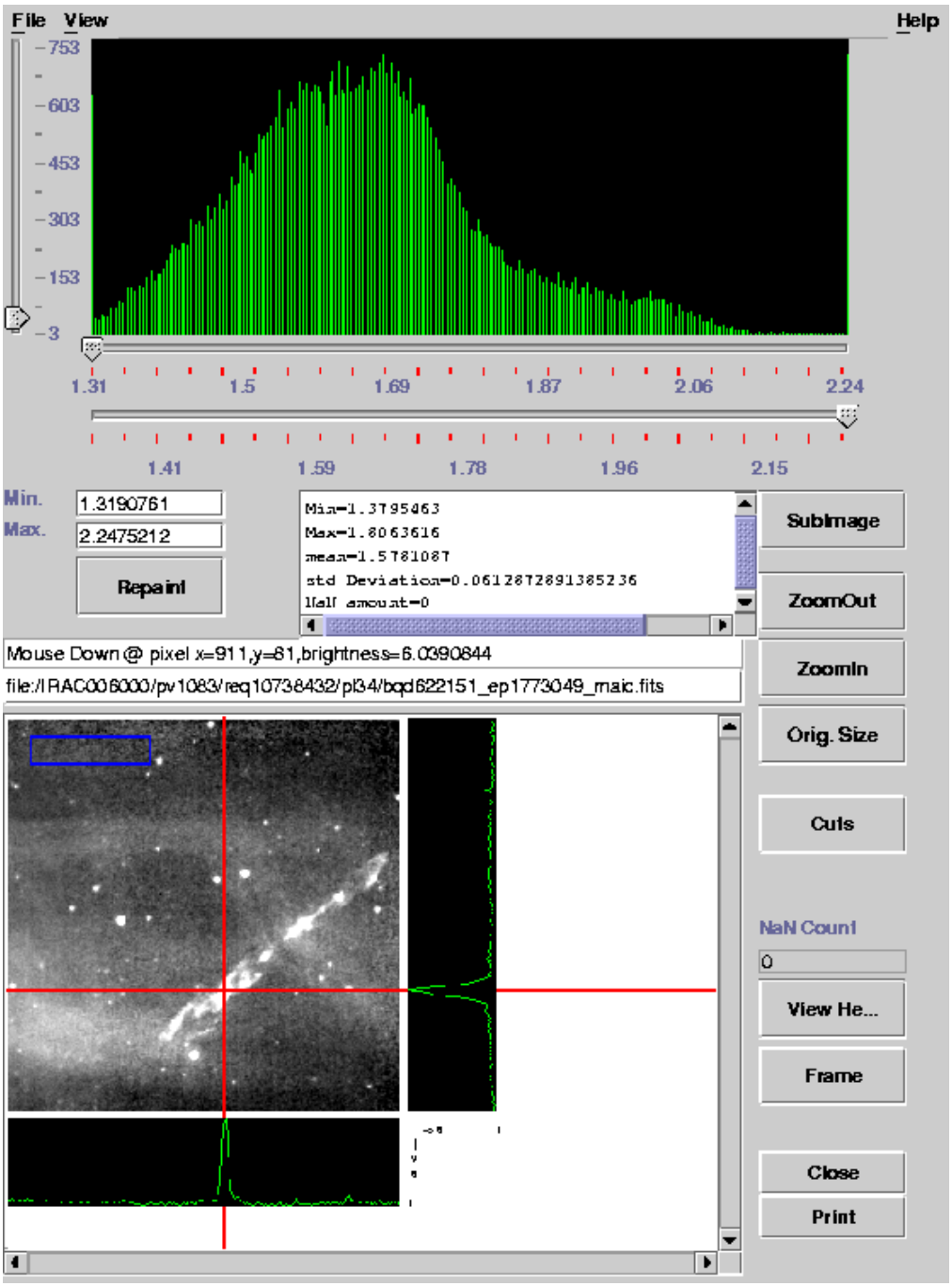


Fig. 5. Screen shot of SDAT, a component of the DQA software system at the Spitzer Science Center. Shown is a channel-3 IRAC map. See Figure 4 for its channel-1 counterpart.

A separate Apache web-server was set up just for running the DQA software system. A secure proxy server was also set up to allow the system to be accessed outside of the firewall, which is very useful to the DQA analysts who occasionally need to work from their homes or other remote locations. For system security, RSA encryption and passwords that change every 60 seconds are employed.

At the SSC facility, the computing hardware used by the DQA analysts and for the DQA system infrastructure are Sun workstations, generally with 500-MHz UltraSPARC-IIe CPUs and at least 1 Gbyte of memory per workstation. The operating system is Solaris version 5.8. The web-based portion of the DQA software, of course, is runnable under most Java-enabled web browsers on common computing platforms.

## 5. CONCLUSIONS

The data quality analysis process at the Spitzer Science Center, supported by a robust DQA software system, has proven itself to be very effective for analyzing and quickly troubleshooting the science data collected each day with the Spitzer Space Telescope. The process requires a relatively small team of just four people. Speedy archive and release of data to the community is ensured by running the DQA process in parallel with automated data processing and archiving operations. The highly efficient pipeline processing and DQA operations employed for Spitzer can be regarded as a model for future space science missions.

## ACKNOWLEDGEMENTS

Besides contributions from the authors, the DQA-software-system design, development and testing efforts were performed by a number of excellent SSC software engineers and scientists, including Bob Narron, John White, Irene Bregman, Sherry Wheelock, Dr. Hua Hu, Dr. Ted Hesselroth, Dr. Deborah Levine. This work was performed at the Spitzer Science Center as part of a mission/project managed by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

## REFERENCES

1. The starting point for a wealth of information on the Spitzer Space Telescope and its science instruments, as well as how to download Spitzer data, is http://ssc.spitzer.caltech.edu
2. Flexible Image Transport System (FITS) is the image format in common use by astronomers.
3. Wen Lee, Russ Laher, John W. Fowler, Frank J. Masci and Mehrdad Moshir, "Caltrans keeps the Spitzer pipelines moving," *Astronomical Data Analysis Software and Systems XIV*, 347, 2005, 594-598 (2005).
4. WCS stands for World Coordinate System.
5. H. Brandenburg, P. Lowrance, R. Laher, J. Surace, and M. Moshir, "Using Perl in basic science and calibration pipelines for Spitzer infrared array camera data", *Astronomical Data Analysis Software and Systems XIV*, 347, 2005, 575-579 (2005).
6. Frank J. Masci, Russ Laher, Fan Fang, John W. Fowler, Wen Lee, Susan Stolovy, Deborah Padgett, and Mehrdad Moshir, "Processing of 24 micron image data at the Spitzer Science Center", *Astronomical Data Analysis Software and Systems XIV*, 347, 2005, 468-472 (2005).
7. Russ Laher and John Rector, "New software for ensemble creation in the Spitzer Space Telescope operations database," *Astronomical Data Analysis Software and Systems XIV*, 347, 2005, 594-598 (2005).