NASA/IPAC Infrared Science Archive

Long Term Preservation of Data Analysis Software at the NASA/IPAC Infrared Science Archive

Harry Teplitz
& the IPAC team

Copyright 2011; government sponsorship acknowledged

Data Analysis Tools

- Required for archival research
- Support non-standard analysis
  - Not everything has one-size-fits-all processing
- Support new methods and/or calibration
- Challenging effort for archives to provide long-term curation
  - Evolving technology
  - Loss of corporate memory
  - New demands for functionality
- This talk: I will describe the multi-decade curation of data tools by IPAC and IRSA
  - starting in the 1980's and continuing into the future

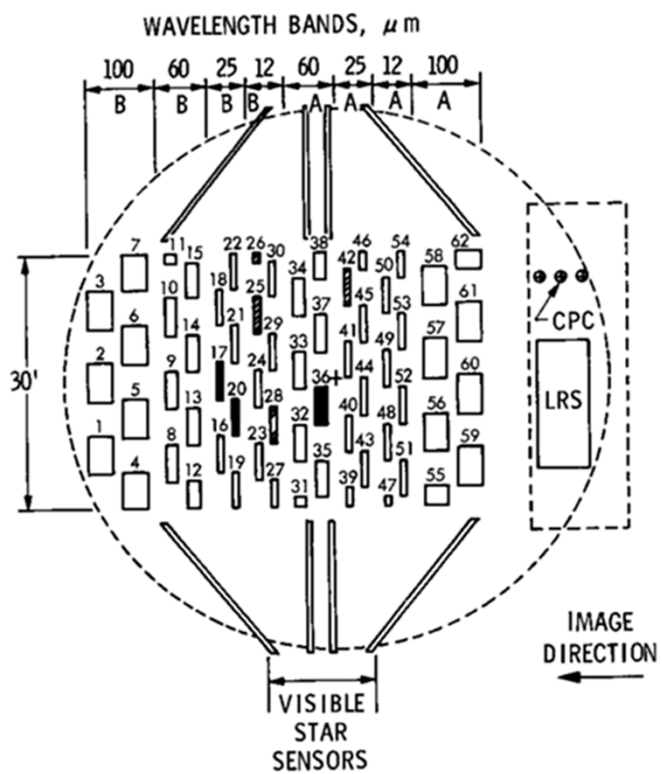
1980's

IRAS

- Mission
  - 11 months in 1983
  - 12,25,60,100 microns
  - 96% of the sky
- Instruments
  - Survey Array
  - Low Resolution Spectrometer
  - Chopped Photometric Channel
- Data archive, and analysis tools continue to be in use today
  - HIRES
  - SCANPI

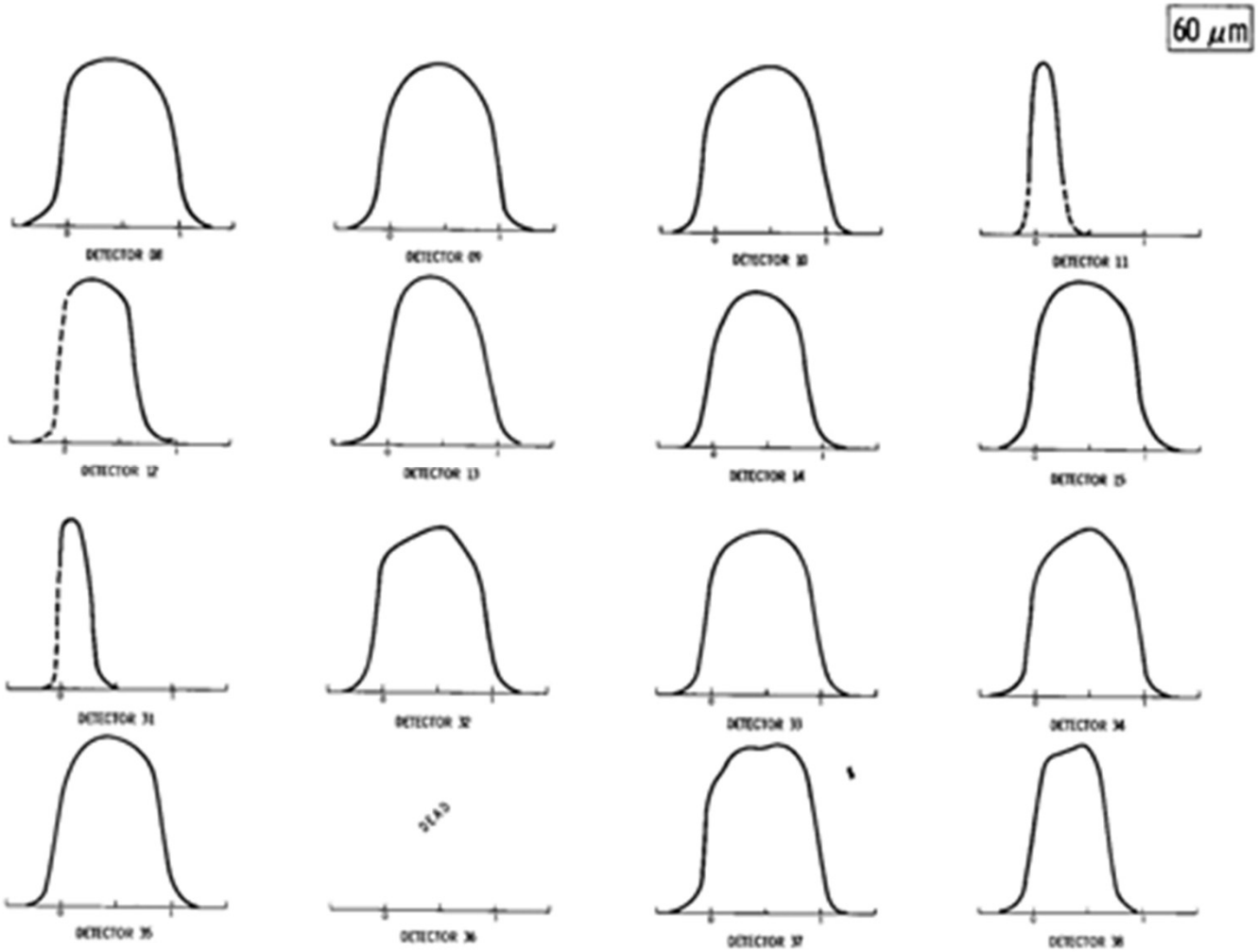
IRAS Survey Data

- IRAS focal plane consisted of apertures leading to integrating cavities
- The detector masks were rectangular in aspect and infrared sources scanned across the focal plane parallel to the narrow dimension of the detectors in all observational modes.
- Measurements reported as Time Ordered Data
  - Objects cross the vertical "array" of sensors;
  - speed of the scan is known;
  - thus shapes can be reconstructed



NASA/IPAC Infrared Science Archive

Cross-scan response of the IRAS 60 micron detectors



Extreme variation of the Point Response Function (PRF)

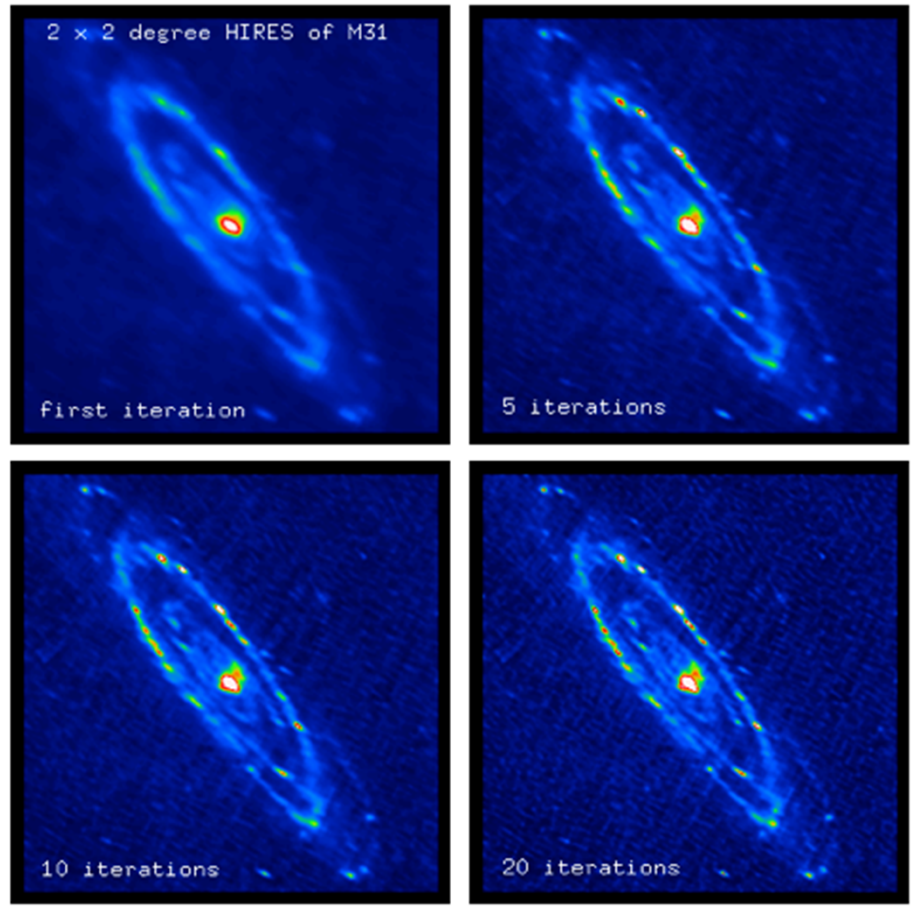
NASA/IPAC Infrared Science Archive

Maximum Correlation Method

- Based on the idea of a simulator that could use a trial image and an observational model to predict what the measurements would be, compare those to actual measurements, and use the differences to feed back a correction to the trial image; this process was then iterated until it converged
- Handles the case that every measurement (sensor or pixel) has a PRF
- Makes use of noise measurements
- Operates on overlapping measurements with, if necessary, completely under-sampled PSFs
- Reduces to Richardson-Lucy algorithm for isooplanatic oversampled PSF with no prior noise estimates
- H.H. Aumann, J.W. Fowler and M. Melnyk, (1990), *AJ*, ***99, 1674***

HIRES

- HIRES uses MCM to produce images with better than the nominal resolution
- Powerful tool for studying morphology and separating confused sources
- Produces IRAS resolution better than an arcminute (5x increase over baseline)
- Fluxes good to 20% (due to background determination uncertainties)
- Available as 1x1 or 2x2 degree images, with 15 or 30 arcsecond pixels
- User controls which scans are used as input, and what de-striping (background and time-varying detector response correction) technique is used



NASA/IPAC Infrared Science Archive

HIRES

- Originally ran on the Cyber computer at IPAC
- Algorithm was designed to be extremely sensitive to nuances in the data, and there were plenty of artifacts and poorly conditioned data that required a human in the loop.
  - IRAS project was originally designed only to produce a point-source catalog,
  - resources could not be allocated to control the electronic baseline (which was automatically subtracted off by the point source detection/extraction algorithm).
  - As a result, the goal of HIRES imaging was initially limited to objects seen against lots of dark sky which could be used to locate a baseline.
  - Extended structure caused problems for baseline registration across detectors and across scans.

SCANPI: The Scan Processing and Integration tool

- Developed from ADDSCAN
- Performs 1-dimensional (in-scan) scan averaging of the IRAS raw survey data
- IRAS raw scan data, aka the Calibrated Reconstructed Detector Data (CRDD) are stored in files merged with the IRAS pointing (boresight) information
  - This Level 1 Archive, and it is stored in compressed format on an archive server maintained at IRSA.
- Scanpi averages the IRAS scans with various weighting schemes. It is used to obtain the fluxes of extended, confused or faint sources, and to estimate true local upper limits.
  - sensitivity gain is comparable to that obtained in the IRAS Faint Source Survey (2-5x over IRAS PSC)
- Fit detector profiles
  - Incorporates detailed knowledge of the instrument and observation strategy
- Extract fluxes

1990's

HIRES & SCANPI Porting

- Ported SCANPI to Sun/Unix from Cyber (1991)
  - Part of the effort to create the Final Archive for calibrated detector data and boresight data
- Ported HIRES to Sun/Unix (1992)
  - Converted QA from contour plots to images
- Storage Media
  - Archive data stored on 8mm tapes
  - Benchmark Tools results also stored
- Stop human QA on HIRES output in ~1995
- Convert documents to electronic form

HIRES algorithm improvements

- FITS header output includes –CAR for projection type (10/5/95)
- Algorithm improvements (4/24/96)
  - more effective methods of removing electronic baseline and "de-striping" were developed (e.g., the IRAS Galactic Plane images; Cao et al. 1997 ApJS, 111, 387),
  - but the nature of the data prevented complete automation; a human looks at the output to check

2000-2011

SCANPI upgrade 2007

- Supported by Spitzer
  - Needed by Spitzer observers for IR research
- Program-friendly interface added
- Significant re-structuring of code
  - Bugs-fixes and other small changes had been performed without sufficient documentation
    - Loss of corporate memory
  - Decades of small fixes led to unnecessarily complicated code

Adapting MCM

- The MCM was adapted from HIRES to use with imaging telescopes (e.g. Spitzer)
- First adapted for WIRE
- Further developed for WISE (e.g. Masci & Fowler 2009)

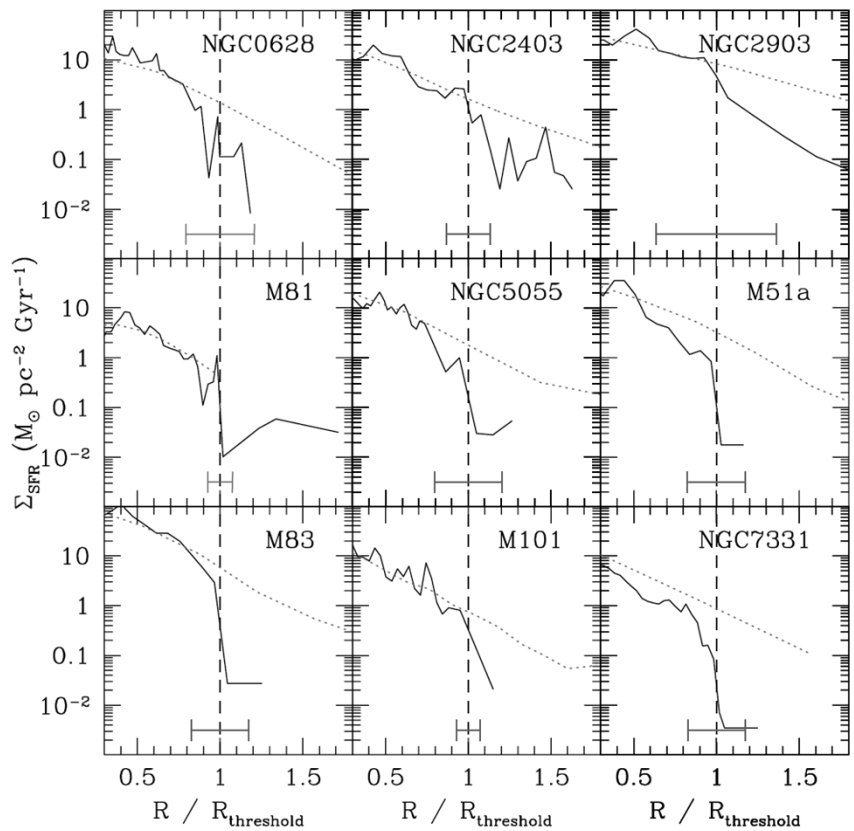
Figure 2.: M51 from *Spitzer*-MIPS 24 μm. *Left*: co-add (after one iteration); *Middle*: HiRes after ten iterations; *Right*: HiRes after 40 iterations.

NASA/IPAC Infrared Science Archive

HIRES at IRSA in 2007

- In 2007, there were 5,050 separate requests for HIRES images
  - Labor-intensive
  - Resources did not permit major update to the code

- HIRES maps used in combination with GALEX to test models of ultraviolet extinction in large galaxies (Boissier et al. 2007)



NASA/IPAC Infrared Science Archive

HIRES at IRSA

- Porting from older Sun computers
  - Crash in 2008 required installation of equipment from the right era
  - Limited access to older hardware
- Re-compile on newer Sun
- Port to Linux, following larger IRSA migration
- Increase the level of automation in processing user requests
- Regression testing
  - Relies on data stored on 8mm tape

Spitzer: major increase in tools to be curated by IRSA

TABLE 6: Spitzer Data Analysis Tools	
Tier 1 (absolutely required for the archive)	
MOPEX/APEX	Mosaicking and source extraction for all Spitzer imaging (IRAC, MIPS, IRS-PeakUp). Enables IRAC photometry at the few percent level, accounting for intra-pixel sensitivity. Standard software (e.g., daophot) is less effective
SPICE	IRS staring-mode spectral extraction; simultaneous visualization of slit-position in 2D dispersed image and a direct image
CUBISM	IRS mapping-mode spectral extraction; construction of data cubes; CUBISM supplies a unique functionality for which there is no accessible alternative
GeRT	The MIPS 70/160 pipeline allowing for custom re-reduction; reduction is particularly sensitive to observing details
CUPID	IRS pipeline; vital option for power users
IRSCLEAN	Cleans bad pixels from 2D IRS spectra
Tier 2 (highly desired but unfunded)	
BANDMERGE—merges catalogs, taking into account positional uncertainties **DARK_SETTLE**—mitigates time-dependent dark current in IRS spectra **IRSFRINGE**—de-fringes IRS spectra **PAHFIT**—fits PAH lines in IRS spectra	

NASA/IPAC Infrared Science Archive

Documentation for Spitzer Tool Curation

- Plan developed with Spitzer Science Center
- User Manuals
- Tutorials/cookbook
- Code overview document and better comments in c-code to ease modification of the source code.
- Build Instructions to explain how to make binaries.
- Regression Testing Plan – Data sets and namelists that fully exercise the code, along with outputs and scripts to find discrepancies wrt previous versions. Complete for MOPEX and SPICE.
- Tool Curation Plan

Science User Support

- It's not enough to have the software and the documentation
- Users expect help with using the tools
- Need to evaluate bug reports
- Tools expertise is an important part of retaining corporate memory

The Future

WISE Custom-Coadder

- MCM adapted from HIRES is used in WISE
  - Standard image construction module in data reduction pipeline
- Pipeline module will be adapted as web-based user tool
  - Enables co-addition based on epoch; mission products can't cover all cases
  - existing WISE image coadder software will be modified to interact with a web interface.
  - Simple web interface will be developed, in the context of existing IRSA UI, including HIRES
  - underlying middleware to manage data exchange with the L1b image archive available as part of other IRSA services

WISE 3-color Atlas image of NGC1514

NASA/IPAC Infrared Science Archive

HIRES and SCANPI for Planck

- Planck is in important ways similar to IRAS, with nearly identical survey strategy and a focal plane filled with heterogeneous, not-diffraction-limited detectors.
- Leverage IRAS experience by adapting HIRES and SCANPI to Planck data
- SCANPI, would extract all scans through a given celestial position; the scans will be individually available, baseline-removed, and coadded.
- HIRES, will take all scans through a rectangular patch of sky and combine them into a finely-gridded mosaic,
  - using a maximum entropy or maximum correlation method to replace the asymmetric Planck beam for each detector with a smaller, circular Gaussian.
- Closely tied to IRSA infrastructure and UI

Conclusions/Lessons Learned

- Retain corporate memory
  - Including Science User Support
- Documentation
- Regression testing
  - Storage of benchmark results (tapes unreadable/degrade)
- Resource Limits on upgrades
  - How do you decide whether the upgrade is worth it or the tool should be retired?
    - Cost, demand, relevance of the data set
- Take advantage of connection to archival infrastructure
- Advantages of generalizing beyond a single mission
- Hard to predict compute improvements over decades