Big Data at the NASA/IPAC Infrared Science Archive (IRSA)

Steve Groom

Caltech

IRSA: NASA/IPAC Infrared Science Archive

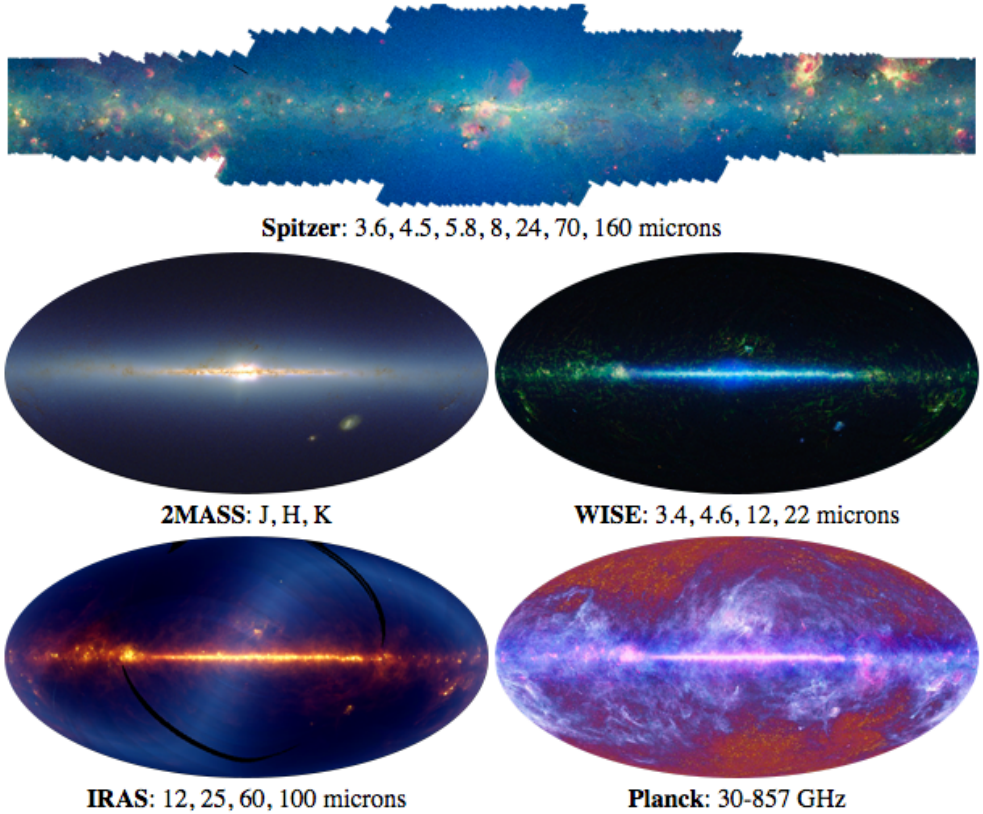
Enabling research that has not yet been envisioned

IRSA's Charter

- IRSA is chartered to curate the science products of NASA's infrared and submillimeter missions.

IRSA's Holdings

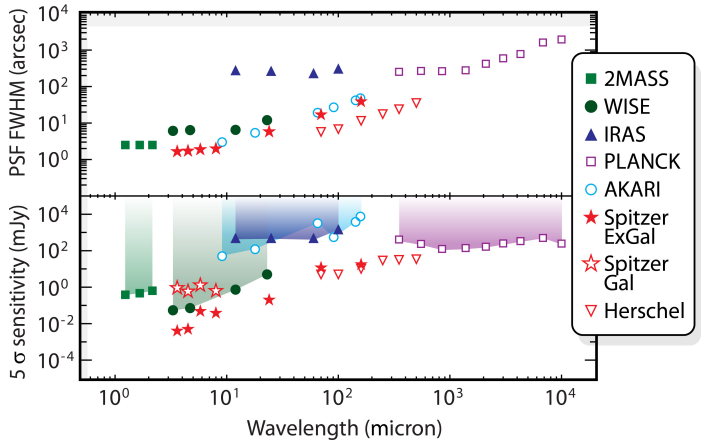
- All sky images and catalogs covering 20 wavelengths.
- Spitzer, NASA's Infrared Great Observatory
- >1 PB since 2014
- >100 billion table rows



IRSA: NASA/IPAC Infrared Science Archive

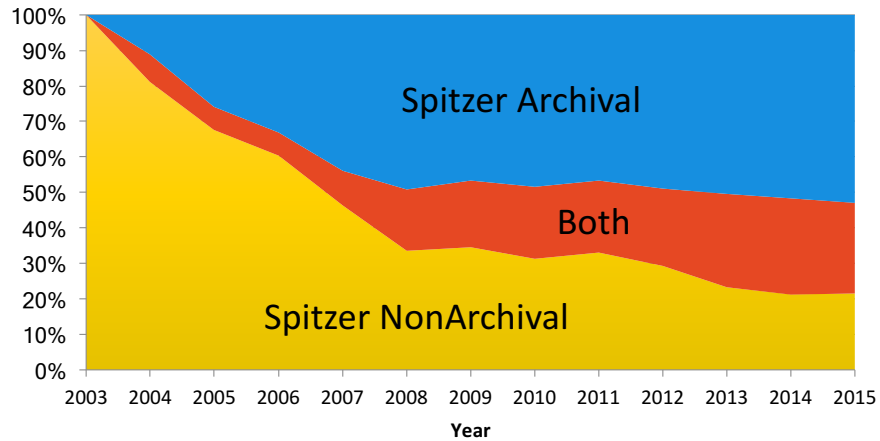
Enabling research that has not yet been envisioned

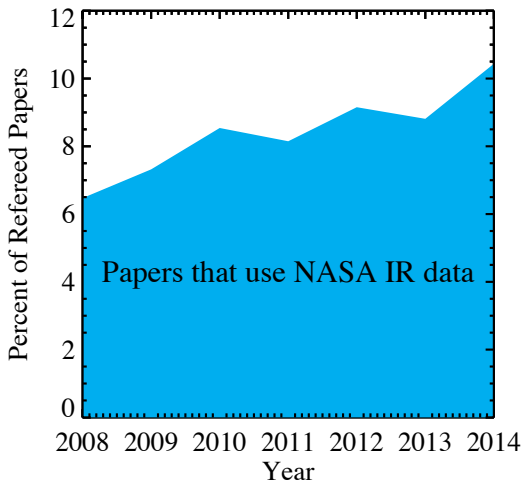
- IRSA's diverse holdings applicable across the range of NASA's Astrophysics themes
- Richer content draws new, more complicated potential use cases
- Holding the bits alone is not sufficient to enable new science
  - Archive provides support to understand what the bits mean and which ones to use
  - Archive assists users with logistics of access and analysis

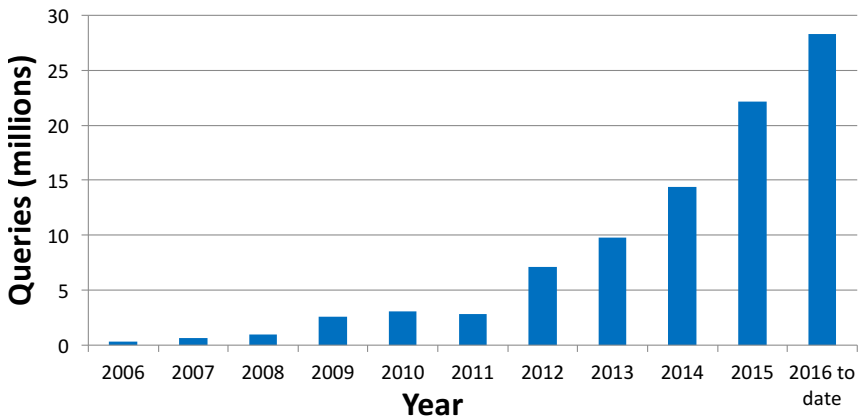


IRSA's Impact

- Archives can double the number of papers from a mission.
- ~10% of refereed astrophysics journal articles use data in IRSA's holdings.
- IRSA over 25 million queries this year to date.







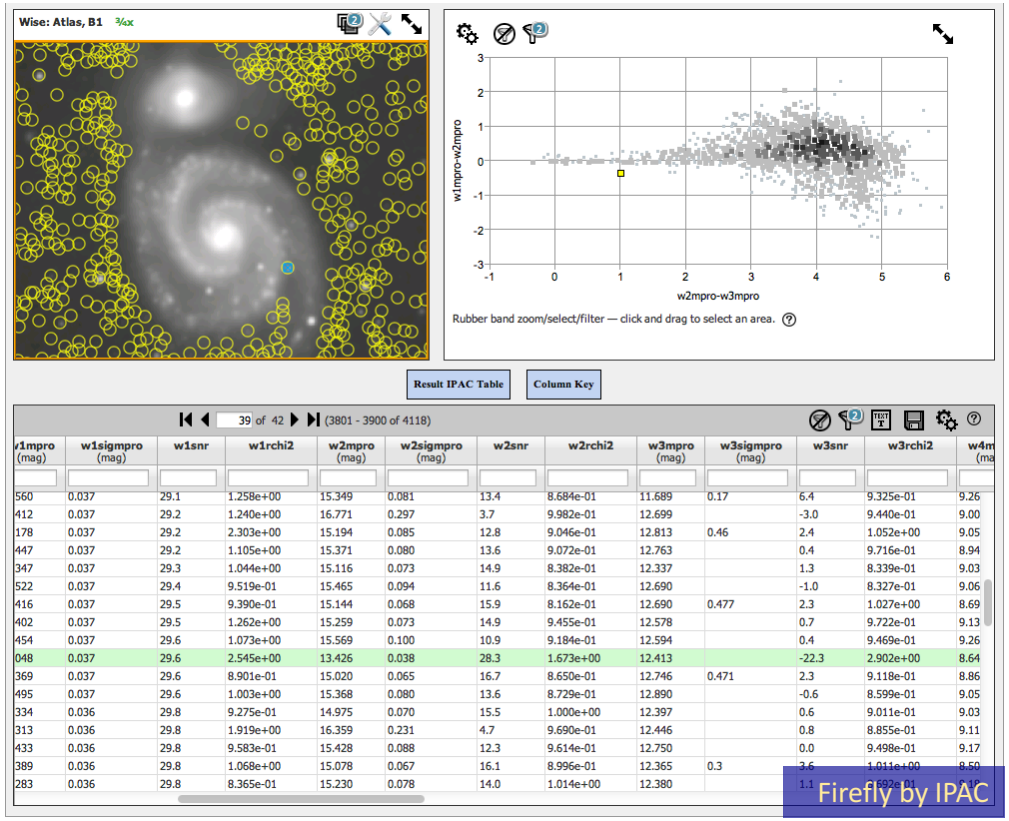
Operational Technology

Large Table Management

- Large tables, growing regularly, while maintaining performance
- HTM indexing used for spatial searches
- Bulk queries use matching outside the database

Visualization: Firefly by IPAC

- Archive-aware interactive web tools
- Linked images, plots, tables
- Reusable components permit both general and project-tailored data views
- Cost effective to maintain and develop
- API supports interoperability

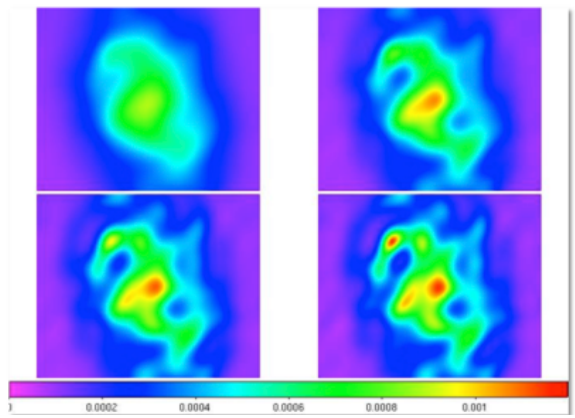


Operational Technology

Data Analysis at the Archive for manipulation of pixel data

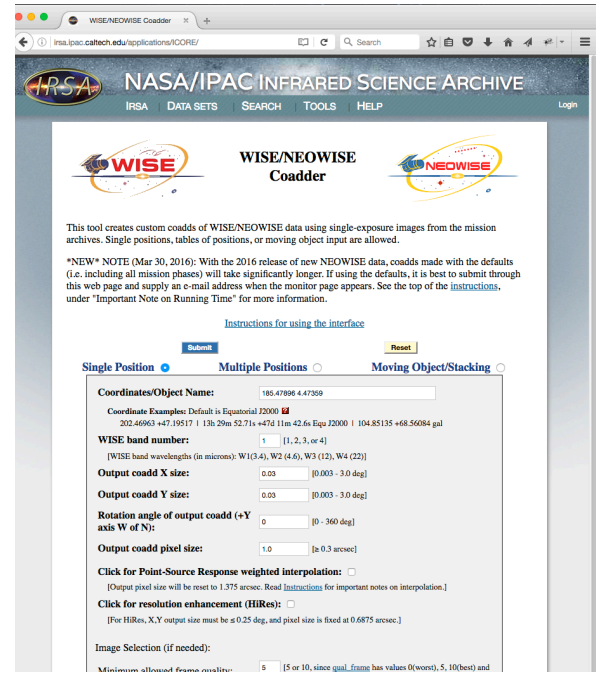
- Image Enhancement: HIRES – IRAS, Planck
- WISE/NEOWISE Custom Coadder

HIRES processing of 857 GHz Planck image of M33



- Python-based re-implementation of Maximum Correlation algorithm (MCM) originally developed for IRAS data (Aumann, Fowler & Melnyk 1990)
- 45x45 arcmin HIRES image of M33 derived from Planck TOD
- shown at iteration counts of 1 (initial map), 5, 10, and 20.
- improvement in PSF is ~2x (final FWHM ~2.3 arcmin at 857 GHz – iterations 5-20)

HIRES implements the Maximum Correlation Method algorithm developed for IRAS (Aumann, Fowler & Melnyk 1990), adapted to perform resolution enhancement on Planck data



WISE Coadder tool allows users to create customized image coadds (stacks) of WISE data directly from archive

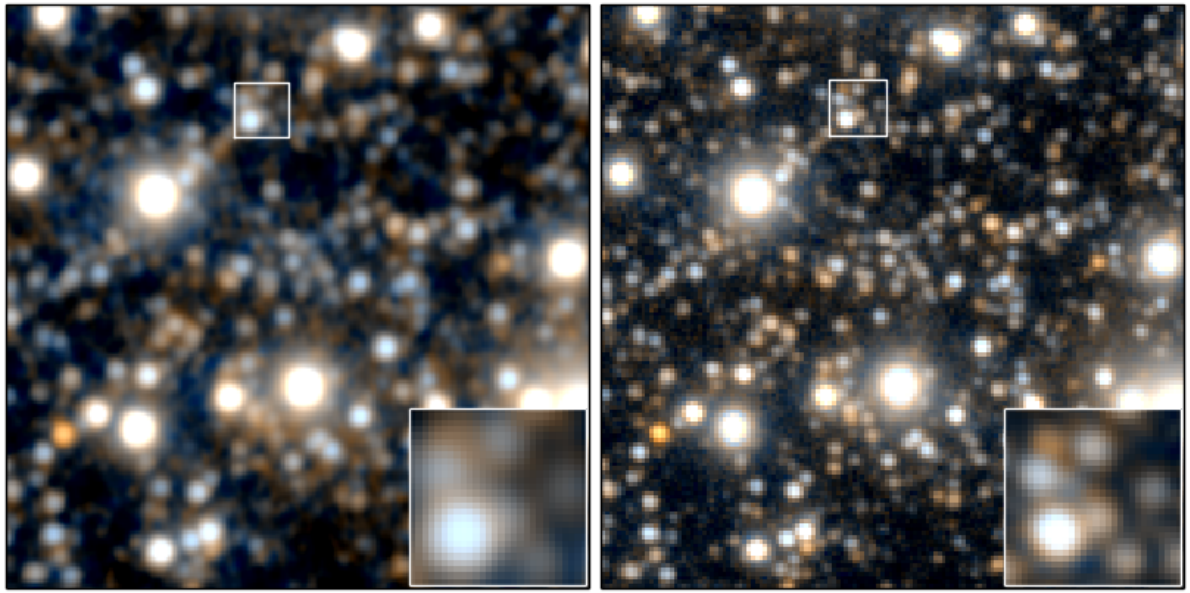
Science Exemplifying Big Data

Large scale reprocessing

- Driven by synergy between an IRSA dataset (WISE) and an external dataset (SDSS)
- Resulting product has potential applications to many questions
- IRSA presently accepts contributed enhanced products

Logistics of Big Data

- Required access to all the pixel data
- Movement of all the data presented a major challenge
- Processing all the data required compute resources beyond what the archive could offer

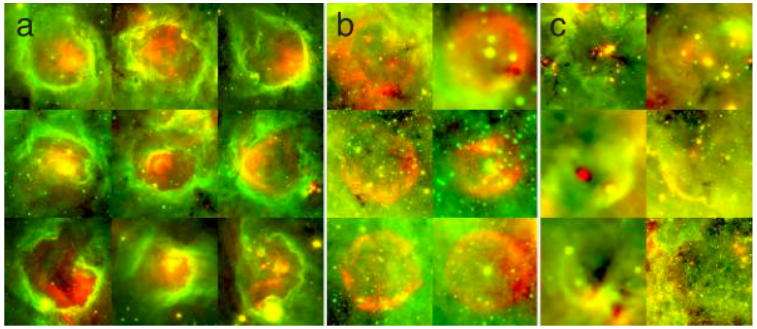


Lang (2014) reprocessed WISE single-exposure images to optimize measurements for extended sources.

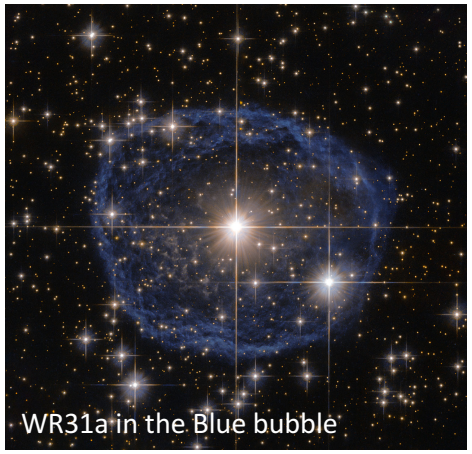
Science Exemplifying Big Data

Finding features in the data

- New interest in machine learning
- Starting to unlock the potential of discovering patterns by looking at all the data at once



Beaumont et al. (2014) compare machine learning and citizen science searches for interstellar bubbles in Spitzer images.



Morello et al. (in prep) are using machine learning on NASA IR catalogs (Spitzer, 2MASS) to identify Wolf Rayet bubbles that are not detectable at visible wavelengths.

Big Data Implications

- Data discovery tools will become increasingly important – selecting the right data
  - Data flooding in from many projects
  - Interoperability means IRSA users have access through the VO to a multitude of data sets
  - How can we help users find the ones they are most interested in?

- Analysis at the archives will become increasingly important
  - Deep analysis and custom processing requires general access to all the data
  - Fixed services won't sustain flexible, growing demand
  - "Big data means you can't move it all", suggesting **the analysis must move to the data**

- Archives need resources to meet science analysis needs
  - Potential for improvements to existing services to support out-of-archive analysis
  - Need resources to stay abreast of technology landscape, important in shaping solutions
  - Technology pilot efforts given lower priority in 2015 Archives Senior Review

The Archive Paradox

- NASA's Astrophysics Archives have focused on curation and online query/access services for science datasets. Resources available for custom processing and analysis are limited.
- Mission proposals primarily focused on generation and distribution of standard products
- Many users have indicated interest in accessing resources for deeper analysis and mining of those products

Incompatible Realities:

- Traditionally, archives provide access to data and some limited, shared resources, while researchers/projects cover their specific analysis resource needs
- Big data means you can't move it all, it must be analyzed in-place

New paradigm needed?

Possible approaches

- Partner with NASA and other potential sponsors on developing a more analysis-friendly model, better supporting in-archive processing

- Use cloud storage and computing resources as a convenient middle ground for shared purposes, some kind of archiving & analysis "co-location" services
  - *Are archive data volumes compatible with cloud cost models?*

Planning for the Future

Panel question #1:

- *What are the processes for planning for future (5—10 years) capabilities of your service?*
- *How and from whom do you gather input for this planning process and where does input typically come from?*
- *What new feature(s) do you really want to implement?*

- Project Priorities
  - data-providing missions, e.g. WISE, Spitzer, IRTF, SOFIA have strict schedules for data releases and closeout.
  - Serving mission data releases take high priority
- User Priorities
  - Twice yearly meetings of our User Panel result in written reports based on presentations.
  - Once every few years, we survey online users.
  - Mission teams provide deep, specialized insight.
  - ADEC, NAVO meet once a month and we hear from our sister archives, giving a picture of community expectations.
- Archive Senior Review
  - Typically considers 4-5 year window
  - Have requested Big Data studies, e.g. cloud computing, as investments in the future, but unfunded.

Retiring Older Technologies

Keeping operations and maintenance under control

Panel question #2:

- *What feature(s) of your service would you like to stop performing?*
- *How do you gather input for making such decisions and where does input typically come from?*
- *What is preventing you from stopping?*

- Some older data sets rely on mission-specific software tools which require maintenance and updates from time to time. Older visualization tools may stop working under newer OS/ browser/language combinations, requiring rewrites.
- These software refresh activities must be incorporated in schedule between high priority activities and development of new functionality
- Tool refresh tasks discussed with User Panel
- Tools can be critical to interpretation of the data

Interoperability

Panel Question #3:

- *What steps are you taking to make your data interoperable with allied data sets from other data sites in and out of NASA?*
- *How do you find allied data sets and what criteria make data sets candidates for enabling interoperability?*

- APIs allow programmatic access to data and tools
  - Archives can query each other, e.g. NAVO, ADEC
  - Desktop tools can query archives, e.g. TopCat, ds9
  - Mission pipelines can query archives, e.g. PTF/Spitzer pipelines can query 2MASS catalog for astrometry
  - All IRSA datasets available through standard API's
- NASA Astrophysics Virtual Observatory (NAVO) ensures NASA archives adopt VO protocols
  - Catalogs: SCS, TAP already implemented
  - Images: SIA already implemented
  - Spectra: SSA in progress
  - VOSpace: in progress
  - NAVO activities managed by NASA, but provide access to and from non-NASA services, as well
- In the era of big data, we will need to co-locate some datasets to enable full interoperability
  - IRSA gets input from user panel, other ADEC members, NASA missions/scientists, and community to identify candidate data sets

Summary

- Increasing data volume and diversity are attracting new, more complex use cases
- Archives confronting challenges responding to science-driven resource demands
- Fuller science exploitation of the archives' potential requires rethinking the traditional scope of archive responsibilities