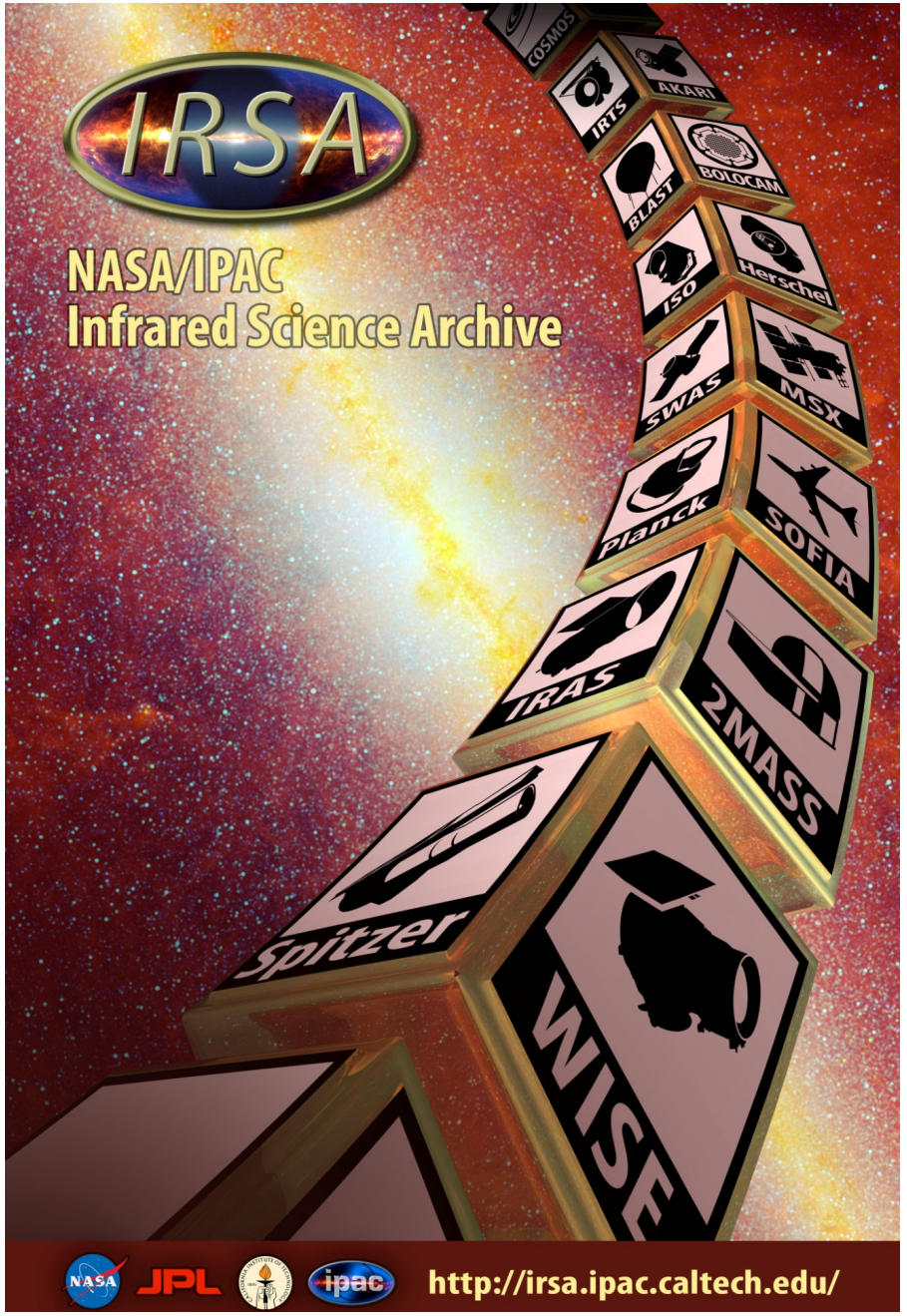
NASA / IPAC Infrared Science Archive

# The IPAC Research Archives

Steve Groom
IPAC / Caltech



# IPAC overview


The Infrared Processing and Analysis Center (IPAC) at Caltech is dedicated to science **operations**, data **archives**, and community **support** for astronomy and solar system science missions, with a historical emphasis on infrared-submillimeter astronomy and exoplanet science.

IPAC is entering its 30th year of building and operating successful data centers and research archives for space- and ground-based astronomical observatories and large-scale survey programs.


# IPAC Activities

- NASA/IPAC Infrared Science Archive (IRSA)
- NASA Extragalactic Database (NED) – *"Google for galaxies"*
- NASA Exoplanet Science Institute (NExScI)
  - Exoplanet Archive
- Keck Observatory Archive (KOA)

IPAC is also home to:

- Spitzer Science Center
- NASA Herschel Science Center
- NASA/US Planck Data Center
- Las Cumbres Observatory Global Telescope (LCOGT) Archive

NASA / IPAC Infrared Science Archive

# IRSA Overview

IRSA is NASA's IR/sub-mm archive

- Began with IRAS and 2MASS, leading into the decade of IR missions

- Spitzer
- Wide-Field Infrared Survey Explorer (WISE)
- NASA Planck Archive


IRSA also provides access to many other mission datasets, including some hosted at other institutions, brought together for the convenience of researchers

Spitzer WISE Planck Herschel 2MASS IRAS AKARI MSX SWAS IRTS BLAST SOFIA

NASA / IPAC Infrared Science Archive

IRSA

# Purposes behind archiving

- Permits others to exploit data – now and into the future

- Provides efficient access to mission data to *enable research that has not yet been envisioned*

The most important questions our users want answered are:

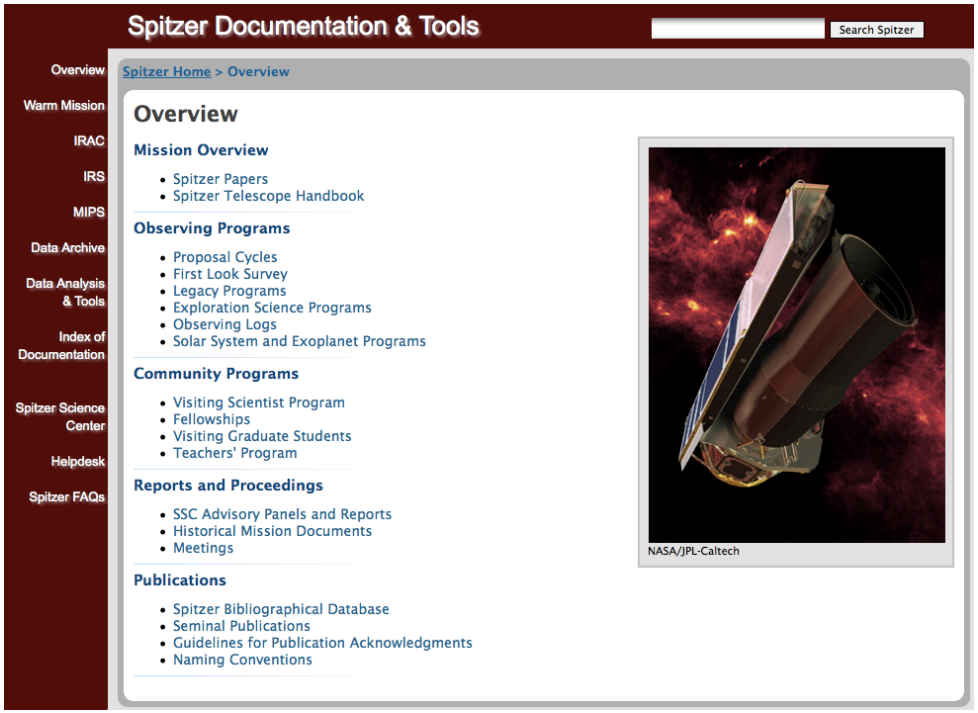
- "*How do I get (and use) my data?*"
- "*Whom do I ask if I have a problem?*"

NASA / IPAC Infrared Science Archive

IRSA

# Archiving involves:

- Data curation
- Data access
- Documentation
- Tools for data reduction, analysis
- User support




All are important to ensure usefulness of the data

NASA / IPAC Infrared Science Archive

IRSA

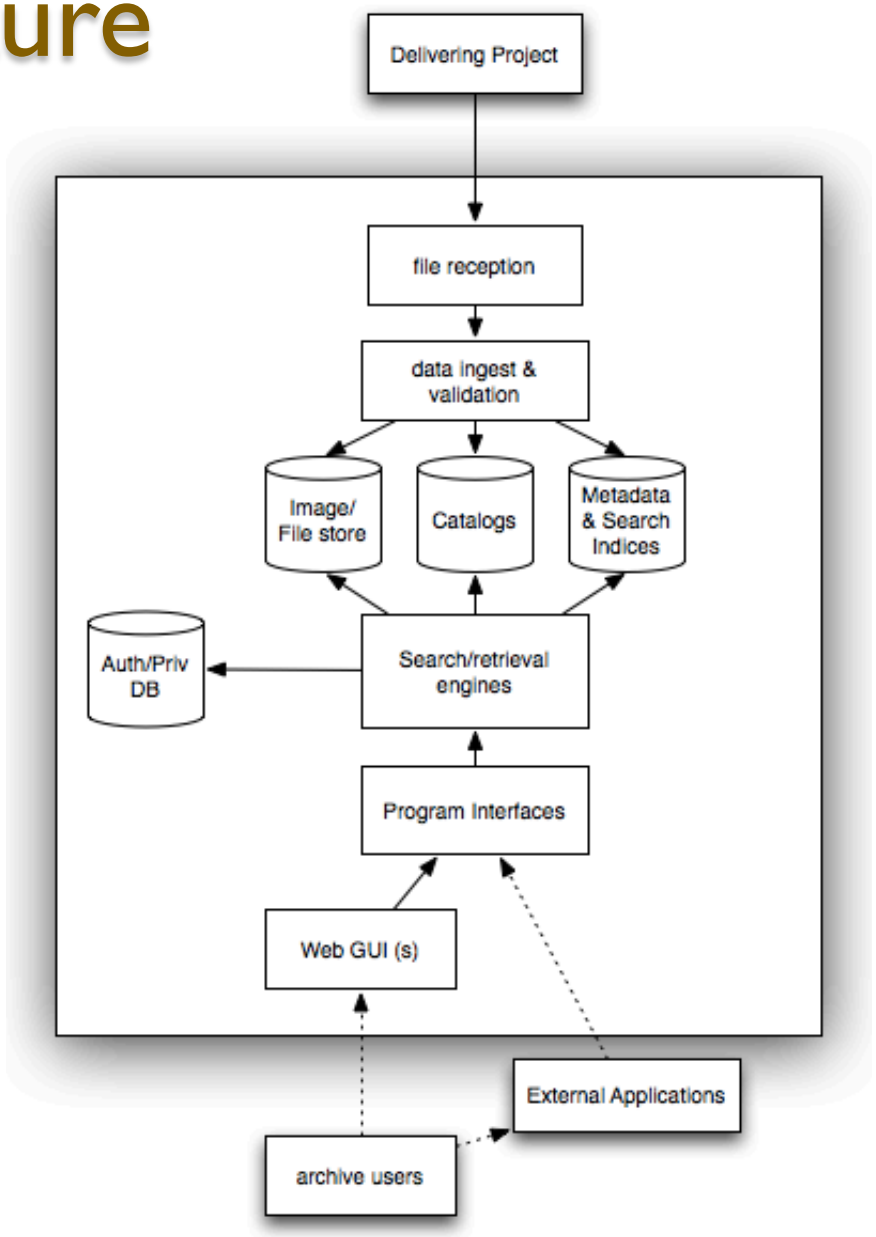
# Data access models

- **Find**: Discovery/Inventory– what's available?
- **Query**: Detailed, dataset-specific query
  - Find records/images matching constraints
- **Combine**: Multi-dataset query
  - Display data from multiple datasets together
- **Mine**: Bulk query
  - Upload list of search positions or other criteria
- **Retrieve**: Bulk dataset download
  - Give me the whole thing
- **Remote Access**: Direct external access
  - "Can I just have my program connect directly to your database?"

NASA / IPAC Infrared Science Archive

# Archive Architecture

- Ingestion & Validation
- Storage, Indexing
- Reusable search engines
  - Catalogs
  - Image/extended spatial
- Product retrieval
  - Also on-demand generation (e.g. image cutout, mosaic)
- Program (VO) interfaces
- Web UI's
- Support for external UI's



NASA / IPAC Infrared Science Archive

# IRSA's common architecture supporting other activities

- NExScI – Exoplanet Archive, Kepler Science Analysis System
- Keck Observatory Archive
  - 20 years of observations
- Solar System – NEOWISE
- Non-NASA:
  - P60, PTF
  - LCOGT
  - LSST
  - TCCON (Atmospheric research)

NExScI
NASA Exoplanet Science Institute

Kepler

LCOGT

TCCON
total carbon column observing network

# Questions to ask in archive design

- What are the products?
  - Standard levels of processing/calibration/reduction?
- What are the use cases for search&retrieval?
- What are the units for data packaging, query, retrieval?
- What can be queried?
  - Metadata – date, instrument, position, etc
  - Pre-calculated summary metrics/statistics
  - Measurement data
  - Derived/calculated properties, combinations of data
- Any proprietary data considerations?
  - Mixing public and private data within the archive can greatly complicate handling of queries involving summary information

# Questions - 2

- Who needs access to the data?
- When do they need access?
  - Real or Near-real time – notification, event distribution: a topic all to itself!
  - Hours/Days/Months/Years later
- Acceptable query latency?
  - Interactive vs batch-style queries
- Anticipated usage/volume

# Things to think about

- Consistency of formats & organization
- Completeness/Correctness of data
  - initial validation is important
- Completeness of documentation
- Examples are helpful!
  - Data reduction "cookbooks"
  - User tutorials, workshops

- Important for the long haul:
  - Consider longevity of technologies, data formats, programming languages/systems
  - Media lifetime, periodic refresh

# Trends: Rapid evolution of information-handling systems

- Technology for data handling changes even more quickly than technologies for photon gathering
  - Fast internet everywhere
  - Extremely portable disks and drives
  - Grid computing and protocols for remote analysis
  - The cloud

NASA / IPAC Infrared Science Archive

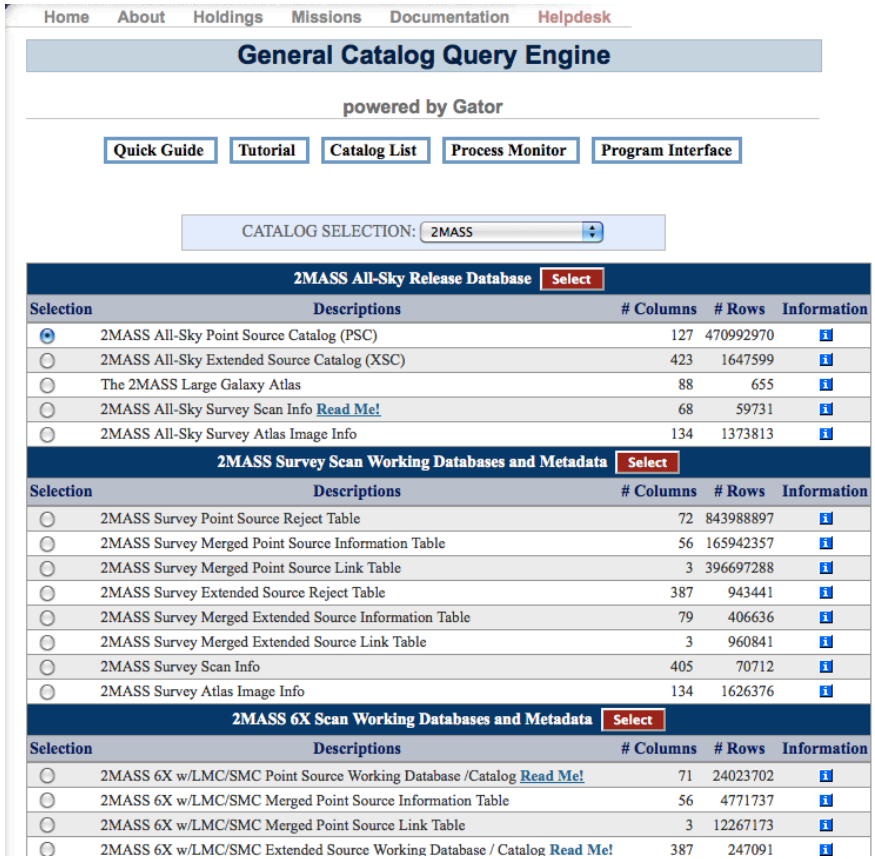
IRSA

# Trends: Archives as analysis environments

- Browse-and-download
  - Identify data of interest, take it home for further study
- Complex queries
  - Finding "interesting" data within large datasets
- In-database analysis?
  - More complex queries over larger data volumes
  - Data size growing faster than communications
  - Bringing the software to the data?
  - *" 'Big Data' means you can't move it."*

# Common Formats and Tools

- Catalogs – tabular data
- Images – usually as FITS files
- Spectra
- Time-series photometry - light curves




NASA / IPAC Infrared Science Archive

# Configurable user interfaces


Decreases the costs of setting up access to new data sets and the costs of long-term maintenance. Provides unified user experience.

# Summary

- Preservation of science data products as well as lower-level products enhances future research
  - Can't predict all potential use cases
- Pay attention to formats, provenance, documentation
  - Future usefulness of the data depends on it
  - "Quality" can be subjective. One man's trash is another man's treasure.
  - Caveat Emptor
- Reusable building blocks can effectively support a wide range of use cases
- Distinct perspectives for rapid vs long-term utilization
  - Fast observation followup vs future data mining
  - Discovery vs research use cases