



The Future of Astronomical Archives

Presentation to the PANEL ON AN ENABLING FOUNDATION FOR
RESEARCH for the DECADEAL SURVEY ON ASTRONOMY AND
ASTROPHYSICS 2020 (ASTRO2020)

Harry Teplitz

Caltech/IPAC

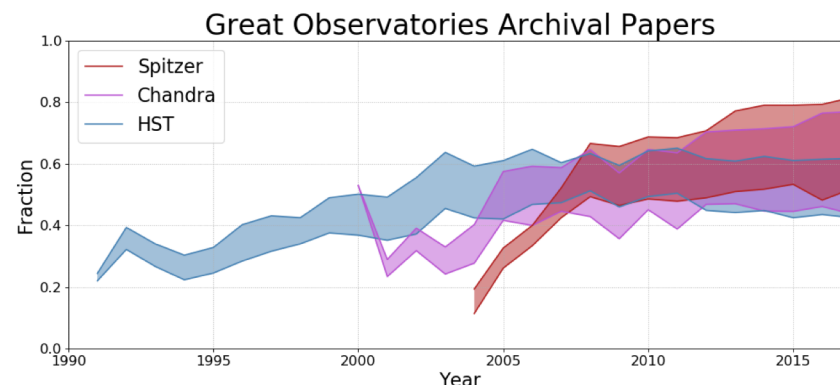
NASA/IPAC Infrared Science Archive (IRSA)

Caltech



NASA's Astronomical Archives Today

- Archival research is a major component of astrophysics today
- Robust funding of archives and research grants has enhanced significantly the science return from major observing facilities and surveys
- Community standards for interoperability are widely implemented and underpin complex services
- Archives preserve not only data, but also mission expertise to enable newly envisioned research
- Archives rely on community input in setting priorities



Fraction of papers using some (lower boundary) or solely (upper boundary) archival data, with no overlap between authors and GO team.

- Archives host all full mission legacy: Raw & processed data; Enhanced science products; Analysis tools; Documentation; Science expertise

Archives double the number of papers from NASA's Great Observatories



Future of Astronomical Archives

- **New capabilities to enhance discovery potential**

- Reprocessing and reproducibility of analysis on large data sets
- Curation of pipelines and software tools
- Integration of community analysis software (e.g. Tractor, R)
- Joint analysis of large data sets needed to extract additional science
- Increasing need for access to simulated observations for direct comparison to data

- **Leveraging technical advancements**

- User-supplied algorithms (e.g. statistical analysis of large data sets) run near the data
- Machine Learning and other AI techniques offer new opportunities for analysis and new tools (e.g. transient identification, SED and light curve classification, cross matching)



5°×2° 18k x 7.2k pixel section (1.2, 3.4 and 8.8 μm) of 16-wavelength Infrared Atlas of Galactic Plane generated from 65TB of data using Montage on the Amazon Cloud.

- **Moving to Cloud solutions as they become financially viable**

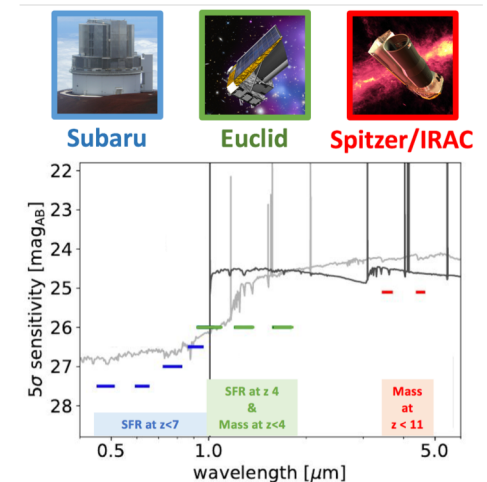
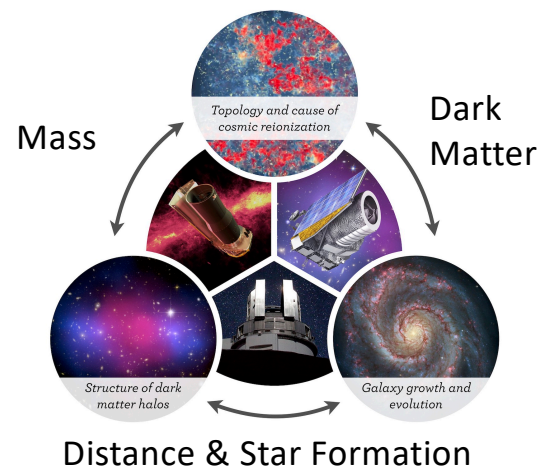
- Ability to scale up and to respond to surges in demand
- Levels the playing field for big data projects
- Agency-level contract or negotiated pricing

Increasingly complex “big data” research will require analysis near the data



Science Platforms Enable Data Processing by Researchers

- Users need tools and environment to explore algorithms and submit large processing jobs.
- Users expect more interactive analysis (e.g. Jupyter notebooks) than currently offered – browsing is not enough
- Requires significant computing resources either at the archive or in the cloud
- High performance networking and/or cloud solutions required, especially in the case of multi-archive analysis



Science Platform = Notebook (e.g. JupyterLab) + Computing resources + APIs + GUIs



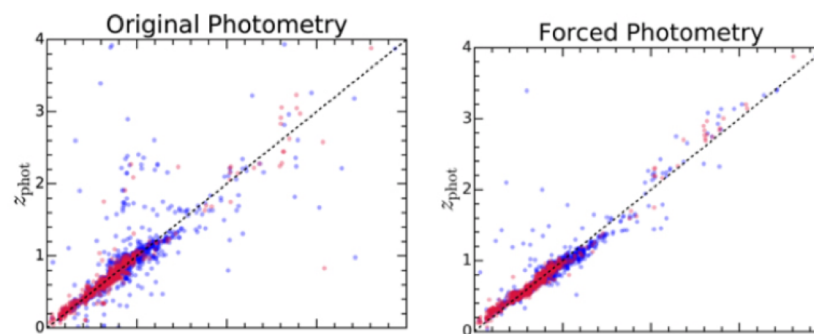
Motivation

- Research is increasingly multi-wavelength
- Joint processing of data from multiple facilities
- Users rely on uniform access protocols to efficiently access multiple data sets

Challenges

- Effective communication between archives required, but resource limited
 - *Solution:* Strong collaborations like NAVO
- Compatibility between internal archive design and interoperability protocols
 - *Solution:* Adopt protocols internally
- Science Platforms could be limited by network speeds
 - *Solution:* High performance and/or cloud-based networking

Interoperability



Improved photo-z in pilot study of Joint Survey
Processing using HST/ACS and Subaru data in COSMOS
for forced photometry with Tractor

NASA and NSF

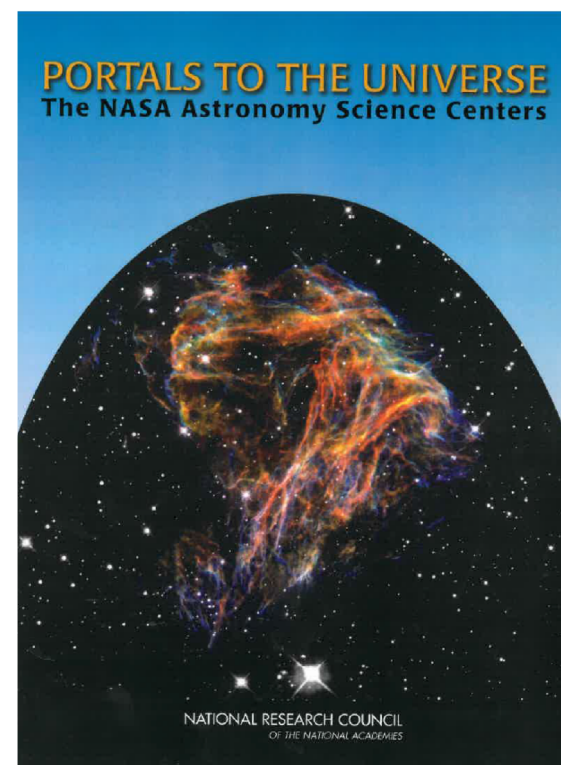
- Different approaches to archival research
- Historical agreement on need for community protocols for data access (i.e. IVOA standards)
- Both converging on need for science platforms
- Interoperability requires sustained resource allocation at both ends



NASA's Sustainable Archives

- NASA supports sustainable archives in an ecosystem that:
 - Has strategic goal to enable more and better science
 - Values and features high-quality, reliable data
 - Facilitates use of community-wide standards for data and services
 - Provides tools and user support to novice as well as power user
 - Has many diverse uses (and users)
- *"Successful research using archival data sets is dependent on the resident expertise and corporate memory that reside at the science centers"*
- Domain-focused archives have the scientific expertise to respond to specific community needs
 - There is no one-size-fits-all solution to archiving astrophysics data: every mission and survey offers new challenges and new science
- Technical solutions shared between disciplines should optimize archive efficiency

Domain-focused archives are the best long-term home for mission data



2007 NRC Report