Discussion Towards Reliable and Productive Astronomical Data Archives

P4-243
(v2020-10-30a)

Hisanori Furusawa[1†], Tadafumi Takata[1], Yuji Shirasaki[1], for data archive workshop participants

1) National Astronomical Observatory of Japan   † furusawa.hisanori@nao.ac.jp

Background and This Study

Astronomical data archives are an essential infrastructure for astronomical activities including cutting-edge sciences. They are also indispensable records of every single moment of the universe – a heritage of humanity. Despite the consensus on the importance, it has been an issue to secure funding for maintaining and improving functionality of the data archives, especially in an era when each community needs to accommodate forthcoming big projects. It is quite important to define solid missions of the data archives and consolidate efforts with proper commitments towards prioritized goals. Specifically in the Japanese community, we have started to investigate demands and optimal ways of operation for reliable and productive Optical-NIR data archives.

To learn from the cases in foreign archives, we have made a private and unofficial survey to several archive scientists who treat raw/processed data, asking how they manage data and what they feel about their circumstances (Table 1; Special thanks to those names listed there). Shown on the far-right column is our case (NAOJ-ADC) for comparison. The tentative outcomes from our ongoing study, for better operation of the data archives, are described below.

Tentative proposal from ongoing discussion

What to archive

- Raw data (that are usable)
- Processed data or at least information of how to analyze raw data
- Quality information on a best-effort basis
- Make as many data available to the public as possible without delay

Major roles of each team

- Data provider (P)
  - deliver raw and processed data to archive
  - validate and fix data for healthy curation
  - perform QA over data
  - prepare tools and/or documentation for data analysis
- Archives (A)
  - preserve data
  - make data available to users
  - enhance values and usability of data

How to decide the datasets

- Data provider (PI, instrument team, or project/mission) should be responsible for decision
- Inputs from relevant community need be solicited for decision
- Archives also need be involved from the designing phase of the instrument and observation, to stably accept data for a long period

Strengths of data archives

- Enhance sample for robust statistics
- Accelerate collecting follow-up data, especially in multi-wavelengths
- Enable detection of rare sources or events with large / long-term sample
- Provide reference for transient events
- Provide evidence for past studies
- Enable possible future new analysis
- Assist education and other social activities

Takeaways from the private survey

In the listed foreign archives that we inquired this time, it seems that they have adequate support from the local community and financial organizations. Roles of groups in the community are well defined. This may be partly because stably continuous and successful achievements with archival data have convinced the community of the importance of the archives, which leads to proper valuation of the archives. We aim to improve Japanese archives, and to make a proposal to our community for reliable and productive future archive operation in Japan.

Question	CADC	ESO Science Archive	IPAC-IRSA	ESA Science Archive	NAOJ-ADC
Q1. What kind of data to archive	All domestic data + some community-involved data (HST, JWST etc) + smaller data sets by request	All raw data from the La Silla Paranal Observatory, and selected internal/external processed data	Most of NASA IR/submm data + all-sky survey data + other valuable PI-based or legacy data	All data from ESA space science missions	All Subaru raw data + other domestic raw data by request
Q2. How much of raw data to preserve	All raw data, indefinitely	All raw data directed, currently indefinitely	All for most missions as directed	Raw data for most of the missions	All raw data requested indefinitely
Q3. How to decide data sets	Archive basically all domestic data	Defined in VLT/VLTI Science Operations Policy	Evaluation by funding agency and grants for big data + Community's opinions for small data	All Science datasets, in agreement with each mission requirements	By commitment with the community for Subaru and NAOJ internal data By request for other data
Q4. Prioritization of data sets	More or less equally treated	Policy mandates that all Large Programmes and Public Surveys return processed data; other datasets prioritized by legacy/community value	Set by NASA, community input, and peer review	Usually, more on pipeline-processed data	Subaru data are more well prepared, but basically equally treated
Q5. Raw or processed data to value more	More on high-level products, but raw data are also equally archived	Preserve all raw data Processed data depends on expected science value, maturity of data processing and operational conditions	Both are equally archived High-level products are more frequently used, but raw data demanded, too	Science data for all missions Raw data for some missions High Level Processed data for some missions	Currently more on raw data? Only limited num of projects provide processed data, and not well established
Q6. Roles of data provider (P) and archiver (A)	P) input validation tests, fix data A) develop & run the validation tests, map metadata to CAOM For some archives, produce advanced data products	P/Phase3 users) deliver data, be responsible for data content and documentation for data releases A) support P, data curation	P) deliver validated data in the proper format with documentation A) ensure pipeline and products conform to archive format assisting P, provide documentation, enhance use of the data	Depends on missions P/P+A) run pipeline and deliver data A) design and develop the archive, in some missions validate data	Depends on projects and no concrete policy P) deliver data A) data curation, validate and sometimes even fix data, on a voluntary basis
Q7. Facility management	Most storage maintained by a separate institute, but still have a copy in-house	In-house. Looking at cloud solutions but not proven to be cost-effective	By other teams at IPAC	By other teams for all ESA missions	In-house. Partly based on a lease contract
Q8. Most important mission	1) maintain a complete collection of all domestic data 2) make it readily available to the international community	Preserve and provide with a long-term time perspective trusted data through science-user oriented services	1) ingest new data 2) maintain the vital IR-data archives 3) enable cutting-edge research	1) provide reliable & validated data 2) keep data for a long term 3) science-oriented and user-friendly services	Shares the same sprit as the other listed archives
Q9. Funding situation	Adequate	Adequate – acceptable through a prioritization process	Adequate for the primary goals – seek additional funding for specific needs	Depends – adequate for funded missions, need to seek funding for legacy-state missions (e.g., Herschel, Planck)	Challenging – manageable only for the existing functions with high priority
Q10. Community's support	Supportive along with the national long-term plan	Supportive – the absolute number and relative fraction of archive publications steadily increase with time, as well as the archive access statistics	Supportive and robust – by the number of publications based on archives	Supportive – improved recently, as most of the publications come out of the archived data	Average – need more robust commitment to maintain and improve archives
Q11. Number of staff	20 FTEs (~8develop, ~4 operation, ~8 research)	Variable -- provided to developments as needed Out-task most operations	12FTEs (~8 develop, ~4 research) for IRSA (HW staff not included). ~1/7 of IPAC budget goes to HW	~30 FTEs for ~25 ESA Space Science Missions	~2-3FTEs per archive (~15 FTEs for ~3 archives and other computing services)

Table 1. Result of survey : courtesy of Stephen Gwyn (CADC), Martino Romaniello (ESO), Harry I. Teplitz (IPAC), and Christophe Arviset (ESA)
We thank them all for fruitful discussion and valuable suggestion on this survey.