Multi-messaging: How can uniform access to astronomical archives be achieved?

Hermann Heßling

[1] *German Center for Astrophysics, Görlitz, Germany;*
hermann.hessling@dzastro.de

Abstract. By combining data from different messengers (electromagnetic radiation, gravitational waves, neutrinos, cosmic rays, ...), one can gain a better understanding of the physics in the universe. A milestone was the merger of a binary neutron star (2017) seen in gravitational waves by LIGO/Virgo, followed by gamma-ray burst, optical/infrared kilonova, X-ray, and radio counterparts.

There are important challenges, for example many messengers are hard to detect, transients often evolve quickly, i.e. a rapid observation of follow-ups is needed, the variety, the velocity, and the volume of data is large in general.

The German Center for Astrophysics (DZA) will provide archives with data from observatories and telescopes around the world. One of the main tasks will be to make it easier for different communities (EM, GW, EM, neutrino, cosmic ray) to cross-match data. This requires standardized data formats, open data, and uniform access, as well as suitable federated data infrastructures

After an overview of the topic, challenges and possible strategies for facilitating access are presented.. The outcome of the discussions with the participants of the Birds of a Feather session (BoF 5) is summarized.

1. Introductory thoughts

The DZA supports the consortium PUNCH4NFDI[1], which is made up of institutions from the fields of astronomy, astroparticle physics, high–energy physics, and nuclear physics. A key objective of PUNCH4NFDI is to contribute the development of a National Research Data Infrastructure (NFDI)[2]. Cloud-based services will be offered, the rough layer architecture of which is shown in Fig. 1. Users can create a container and run it in a remote data center (see the green arrows in Fig. 1) that provides access to storage space (where large volumes of data are archived) and computing hardware.

The DZA aims to support multi–messaging. The cloud architecture shown in Fig. 1 is therefore to be expanded in several respects:

- Astronomical archives offer individual interfaces for accessing their data, i.e., these interfaces should be modified so that access to archives is standardized, thereby making it easier to work with them.

[1]https://www.punch4nfdi.de

[2]https://www.nfdi.de/?lang=en

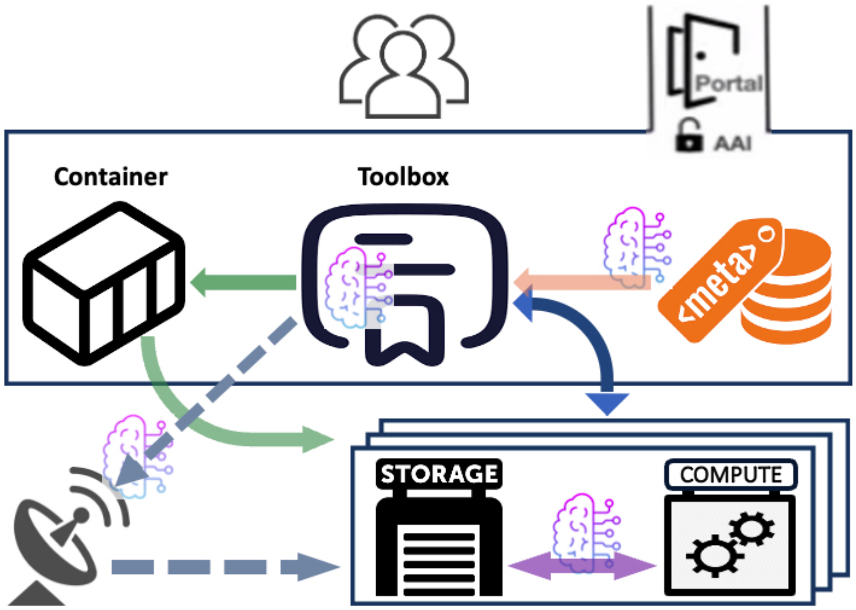

Figure 1. Layered architecture of a federated cloud infrastructure. Illustration by the author.

- A federated AAI (Authentication and Authorization Infrastructure) should be provided to facilitate the management of scientists worldwide who perform analyses on the archive data (see the portal logo in Fig. 1).

- Access to computing power and storage space should be regulated by an accounting system.

- Physicists and astronomers often develop their codes interactively using tools like Jupyter notebook. Therefore, the user should be able to start a container in a remote data center and obtain an interactive connection (see the two–sided blue arrow in Fig. 1).

- Metadata is relevant when data needs to be specified for analysis, and it becomes even more important when it comes to implementing multi–messaging. It is expected that the amount of metadata will grow significantly in the future. As a result, machine learning methods should be available to support the selection of suitable metadata (see the orange arrow in Fig. 1 and the logo with the “neural network brain").

- Most data is stored on tape, and a certain portion is transferred to hard drives, where it can be used for analysis. However, it takes some time to move data from tape to disc. It is therefore advantageous to transfer data in advance, before it is actually needed (Zimmermann et al. 2018). Suitable machine learning algorithms are to be developed for this purpose (see in Fig. 1 and in Fig. 2 the two–sided purple arrow and the logo with the “neural network brain" above it).

- New data from observatories will constantly expand the archives (see the dashed blue lines in Fig. 1). When selecting the data to be archived, it will become increasingly important to develop machine learning methods (see the logo with the “neural network brain" above the telescope logo).

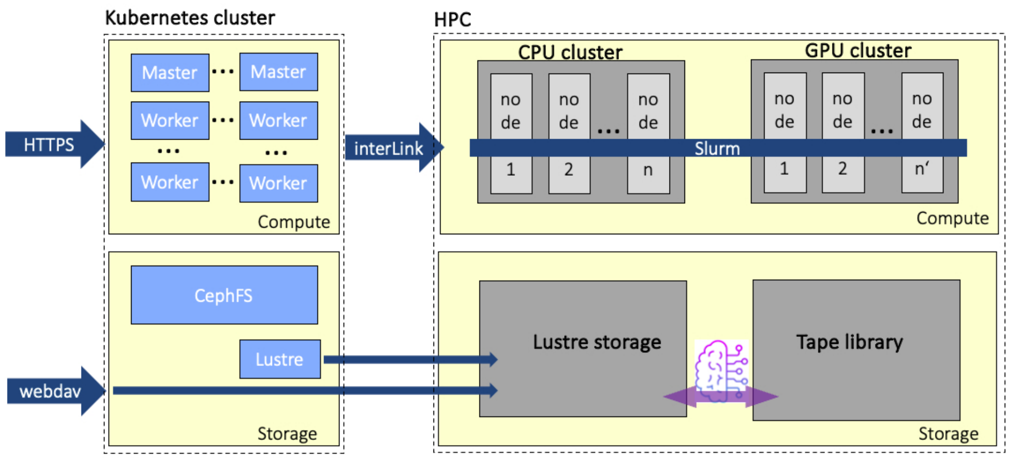

Figure 2. Overview of the architecture of the DZA’s computing hardware. Access to the CPU and GPU cluster (temporarily located in the High–Performance Computing (HPC) center of the Dresden University of Technology during the first years) is provided via the Slurm job scheduling system. Data is stored on tapes or Lustre storage. In addition, a Kubernetes cluster is set up where containers can be started and forwarded to the HPC system using a tool such as interLink. For details, see the poster presentation “Hybrid Cloud & HPC Infrastructure @ DZA" (Borra et al. 2025). Illustration by the author.

2. Important aspects, selected challenges

The Zwicky Transient Facility (ZTF) produces 1.4 TB (compressed) of image data per night and has generated a total of approximately 7 PB of data over the past six years. More than one million exposures have been taken and a total sky area of about 47 million square degrees has been covered, which is roughly 1.5−2 times the total sky coverage expected from the Rubin Observatory LSST over 10 years. ZTF also produces up to one million transient alerts per night, and over 800 million alerts (about 64 TB of data) have been published so far.

Access to data is often facilitated through Jupyter notebooks, which provide a user–friendly entry point for new users. However, resources are finite, even at large institutions, and the diversity of data products – catalogs, images, spectra, and data cubes – raises practical challenges. Decisions regarding the limitation of data downloads appear to be necessary. Storing data in the cloud and accessing it efficiently is costly, and it is unclear whether switching cloud providers will even be possible in the long term, especially from an economic perspective. Additional complexity arises from the need for robust composition, orchestration, and logging of workflows.

Looking ahead, there is a clear need to better support wide-area statistical analyses and large-scale, bulk data processing for machine learning, cross-matching, and recommendation searches. At the same time, the current landscape is fragmented, with proprietary authentication systems and limited ability to share data or computing resources across platforms.

3. Summary of discussion

There was a lively discussion among the participants of this BoF session. The most important proposals and aspects to be considered in the implementation of this BoF are summarized below.

Implementing multi–messaging for federated archives is a challenging task, especially when it comes to realtime reactions to online alerts. It will be necessary to use suitable AAI solutions to enable a standard access to different archives. The AAI architecture developed by the EOSC Association could be of interest here. Given limited computing and storage resources, access must be controlled by means of accounting.

Due to the strong increase in data volumes in the future, processing should take place close to the data (code–to–the–data paradigm). Containerized workflows simplify analysis on remote data and ensure portability. CANFAR, presented at the ADASS conference, could be a suitable platform.

When developing a uniform interface for accessing different archives, taking the user–perspective into account is most important. User-driven development is widely used in software development, meaning that software development involves end users in every stage of the development process. It is hoped that the complexity of accessing archives and ensuring uniform access can be concealed by suitable middleware. Appropriate support/hotline should be provided for users with different technical backgrounds.

It was pointed out that many individual solutions for accessing specific archives were presented at the ADASS conference, which do roughly the same thing, albeit in different ways. The idea came up to bring these solutions together in order to take a first step toward standardizing access to archives. VO (Virtual Observatory) and TAP (Transient Astrophysics Probe) provide widely used standards that ensure that astronomical data from different telescopes, surveys, and archives can work together.

A working group should be set up to further develop standards, submit proposals for their implementation, and collect open questions.

Acknowledgments. We would like to thank Matthew J. Graham for his support in preparing and moderating the BoF, and Elsa Buchholz for her notes documenting the contributions during the discussion.

References

Borra, S., Buchholz, E., Canbolat, U., Desing Raja, C., Drobek, M., & Haupt, L. 2025, in ADASS XXXV, edited by K. Polsterer, S. Wagner, & et al. (Görlitz: ASP), vol. TBD of ASP Conf. Ser., TBD

Zimmermann, N., Wochinger, T., Fuchs, H., Menzel, M., Nagel, D., Thom, F., Strutz, M., & Hessling, H. 2018, Deep learning for predicting the popularity of datasets. https://indico.egi.eu/event/3973/contributions/9333/