Exploring the Universe with Artificial Intelligence (UniversAI)
Proceedings IAU Symposium No. 397, 2026
I. Liodakis & V. Efthymiou, eds.
doi:10.1017/S1743921325104808

Learning from the present for the future: the Jülich LOFAR Long-term Archive

H. Stiele, C. Manzano, A. Miskolczi, V. Vybornov, T. Fieseler and S. Pfalzner

Jülich Supercopmuting Centre, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52425 Jülich, Germany
emails: h.stiele@fz-juelich.de, c.manzano@fz-juelich.de, a.miskolcczi, v.vybornov@fz-juelich.de, t.fieseler@fz-juelich.de, s.pfalzner@fz-juelich.de

Abstract. Since 2013, Forschungszentrum Jülich has hosted the German segment of the LOFAR long-term archive, with current storage amounting to approximately 22 petabytes (PB), exhibiting an annual growth rate of approximately 2 PB. The advent of next-generation radio telescopes is anticipated to generate data at a substantially higher rate, thereby introducing novel challenges in the domains of data processing and storage. Here we present a concise overview of the current data management of the Jülich LOFAR Data Archive, encompassing the ingestion, storage system, export to the long-term archive, and request chain. A detailed analysis of data access over the past decade has been conducted, along with an estimation of energy consumption. The study has identified hardware-related constraints and software aspects as the predominant data storage bottlenecks. From the analysis, we derived requirements for the future expansion of long-term data storage capacity.

Keywords. Radio astronomy, Data archiving, Huge datasets

1. Introduction

The Low-Frequency Array (LOFAR; van Haarlem et al. 2013) is a radio interferometer consisting of a vast array of omnidirectional radio antennas spread across many European countries. Its main hub is located in the Netherlands. LOFAR has an innovative computer and network infrastructure that can handle huge volumes of data. Designed and built by ASTRON in the Netherlands, LOFAR's data are pre-processed on a computer cluster in Groningen, before being distributed to one of the three sites currently forming the LOFAR Long Term Archive (LTA; Renting and Holties 2011): the Samenwerkende Universitaire Reken Faciliteiten (SURF; Cooperating University Computing Facilities) in the Netherlands, the Jülich Supercomputing Centre (JSC) in Germany, and the Poznan Supercomputing and Networking Center (PSNC) in Poland.

Roughly 7 PB of data are ingested into the LTA each year, of which approximately 2 PB are ingested at the Jülich LTA. This data is registered in a catalogue based on the Astro-WISE system (Valentijn and Belikov 2009; Begeman et al. 2012) and is accessible via a web interface†. The Jülich LOFAR LTA consists of a distributed heterogeneous storage system based on the dCache software, which allows huge amounts of data to be stored and retrieved from many different server nodes under a single virtual file system tree with various standard access methods. This hierarchical storage solution comprises spinning disks for "online" or "hot" data and tapes for "offline" or "cold" data. Data ingested into the archive is first copied to the disk pools (cache), allowing for fast ingestion and

† https://lta.lofar.eu/

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1743921325104808 Published online by Cambridge University Press

maximising bandwidth use. It is then asynchronously migrated to tape in a background process.

More details on the data management in the LTA can be found in Manzano et al. (2024). In this paper we also present a data access statistic together with a data access pattern analysis, as well as an infrastructure analysis, including energy consumption. Here we will focus on providing a short overview about the bottlenecks we identified with respect to storing and providing data as well as our thoughts about the next steps in long-term archiving.

2. Bottlenecks for storing and providing data

Data storage is limited by various factors, including hardware, software, and human error, all of which can affect the efficiency of data ingestion and retrieval.

In terms of hardware-related limiting factors, the speed at which data can be ingested is constrained by network bandwidth limitations. Higher network bandwidth enables faster data transfer and ingestion. Furthermore, the size and write speed of the cache pools receiving the data are crucial for efficiently migrating 'cold' data to back-end storage and making space for incoming data. Although hardware maintenance is essential for ensuring the system functions smoothly and downtime can be planned for, unforeseen hardware failures can still impact cache space availability during peak ingestion periods. Implementing strategies such as replicating data across geographically dispersed nodes or using fault-tolerant architectures can minimise downtime and ensure continuous access to cache space, even during maintenance or hardware failures.

Although the software is thoroughly tested in advance, new software versions may sometimes introduce unexpected side effects. These bugs may significantly impact the performance of data ingestion. To mitigate the risk of introducing bugs, it is essential to maintain a robust integration environment and implement continuous integration/continuous deployment (CI/CD) techniques.

Finally, misconfigurations, miscalculations and errors when handling the system can affect the efficiency with which data is ingested. Human errors, whether during configuration or calculation, can lead to sub-optimal performance or even system failure. Therefore, well-documented and well-communicated procedures are essential to minimise the occurrence of such errors and provide robust error-handling mechanisms.

In addition to the above factors, which can negatively impact an efficient data workflow, the data processing schedule can also cause avoidable delays. As the current policy only guarantees that data is stored in the disk cache for one week, a request made after this time has expired may result in data being retrieved from tape, which is far more expensive in terms of access time.

Staging data from tape requires tape drives, which mount and read data from magnetic volumes. The availability and allocation of tape drives is critical to providing timely access to data. The limited number of tape drives, combined with their shared usage across different services, can potentially delay data retrieval and staging processes.

A second bottleneck in the staging of data arises when fewer tapes than pools are used. One tape can only ever be used by one pool. If the files required for staging are on one tape, only one pool can retrieve them. Therefore, if the number of requested tapes is lower than the number of pools, not all pools will actively retrieve data from tape storage. In these cases, the staging throughput is limited by the number of pools and the network throughput to each pool. This issue could be mitigated by using multiple tape replicas and software that can select a replica from which to stage files if the primary replica is in use.

https://doi.org/10.1017/S1743921325104808 Published online by Cambridge University Press

3. Next steps in long-term archiving

Our many years of experience operating the LTA, which JSC began setting up in 2009, will be extremely valuable in meeting the upcoming data challenge. The current setup provides a solid foundation for further developing all aspects of data storage. Based on our experience, we make the following suggestions for future archiving:

Hardware

- A trade-off may be required between scaling up the system – for example, by enhancing the network bandwidth and cache pool performance – and scaling out, for instance, by increasing the capacity to retain more data in the cache without rapidly evicting it after ingestion into the archive. As the amount of data increases, it becomes essential to store the data in the most energy-efficient way possible.

Software

- As the amount of data increases, it is not only the hardware that needs to be adapted to handle larger volumes, but the software too. This includes analysis software, as well as software related to the pre-processing and processing of data. Another challenge is ensuring that this software is as energy-efficient as possible.
- Developing algorithms that can extract Smart Data – data sets extracted from large amounts of data using algorithms according to specific structures that contain meaningful information – from the extreme amount of data will allow us to identify on the fly the data worth keeping. This will offer us the possibility of keeping the overwhelmingly large amount of data we will no longer be able to store entirely at bay. Therefore, the community needs to investigate new technologies and methods, including artificial intelligence and machine learning.
- Another aspect is "Code-to-Data", meaning that the code used to reduce and analyse the data should be taken to the data storage sites and run on the computational resources provided there (i.e. a federated infrastructure), rather than downloading the data to process it locally. Since the analysis software and final data products are much smaller than the data sets analysed, they can be transferred more easily, reducing the amount of data traffic and network bandwidth needed.

Storage of final data products

- Future archiving poses several challenges, one of which is ensuring the long-term availability of final data products. The dCache software is currently flexible enough to allow for the logical and physical separation of data, as well as the ability to treat it differently. Although this is not the primary purpose of the LTA, its data policies could be reconsidered to accommodate the addition of post-processed products to the archive. For the upcoming LOFAR 2.0, LOFAR ERIC intends to indefinitely archive only advanced data products, while raw and pre-processed data will be stored for a limited amount of time†. In the era of multi-wavelength and multi-messenger astronomy, providing simplified access to data products for colleagues who are not experts in low-frequency radio astronomy is imperative.
- To ensure proper handling of research data such as final data products, clear and enforced guidelines for Research Data Management (RDM) are essential. Key aspects include organisation, documentation, storage, security, data sharing, collaborations,

† https://www.lofar.eu/wp-content/uploads/2023/06/Data_Management_Capabilities.pdf

https://doi.org/10.1017/S1743921325104808 Published online by Cambridge University Press

ethical and legal considerations, data preservation, and the implementation of a Data Management Plan (DMP)‡.

Providing FAIR data

• Making data products adhering to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al. 2016) facilitates the scientific progress by enabling the reproducibility, replication, and reusability of research results. To find and reuse data it is indispensable to enrich them with meaningful metadata. Without well documented observing conditions and reduction steps data sets are of very limited use to other users.

• To ensure interoperability, it is essential to use common data formats and standard protocols, and to make standardised software available for reducing and analysing data. The International Virtual Observatory Alliance (IVOA) is responsible for defining standards for data (and metadata) in astronomy. In light of the 'Code-to-Data" approach, it would be beneficial to establish software repositories that offer optimised data reduction and analysis tools for specific machines. This would also help to ensure data interoperability and facilitate its reuse. However, this approach requires either a central institution to provide the tools, or the community to be willing to share code.

Preservation of data

• In the FAIR context a long-term archive should not merely store data; it should preserve it. Achieving this requires a proper funding model (comprising operation and support) to facilitate preservation and ensure it remains unaffected by political decisions or shifts in strategic direction.

• The correct use of metadata, which enhances the meaning of the data and enables assessment and processing by future users, independent of those who initially created it, is indispensable. In the field of observational astronomy, recommendations on the definition and usage of metadata (including metadata standards) are provided by the IVOA and by the Common Archive Observations Model (CAOM).

Environmental impact

• Innovative ideas like feedback on energy usage and carbon footprints are being discussed. Breakthrough concepts are being developed for allocating computing resources, such as applying for watt-hours instead of core-hours. Currently being discussed for computing time allocation, this concept has potential to be expanded to other disciplines, including long-term archiving. Further work is needed, including the definition of meaningful units.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460248186 (PUNCH4NFDI) and by the project "NRW-Cluster for data intensive radio astronomy: Big Bang to Big Data (B^3D)" funded through the programme "Profilbildung 2020", an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia.

References

Begeman, K., Belikov, A., Boxhoorn, D., Dijkstra, F., Holties, H., Renting, G., Vermaas, N., & Vriend, W. 2012, Scaling astro-wise to lofar long term archive. *Experimental Astronomy*, 35.

‡ https://www.fz-juelich.de/en/zb/open-science/research-data-management

https://doi.org/10.1017/S1743921325104808 Published online by Cambridge University Press

Manzano, C., Miskolczi, A., Stiele, H., Vybornov, V., Fieseler, T., & Pfalzner, S. 2024, Learning from the present for the future: The Jülich LOFAR Long-term Archive. *Astronomy and Computing*, 48, 100835.

Renting, G. A. & Holties, H. A. LOFAR Long Term Archive. In Evans, I. N., Accomazzi, A., Mink, D. J., & Rots, A. H., editors, *Astronomical Data Analysis Software and Systems XX* 2011, volume 442 of *Astronomical Society of the Pacific Conference Series*, 49.

Valentijn, E. & Belikov, A. N. 2009, Lofar information system design. *MemSAIt*, 80, 509.

van Haarlem, M. P., et al., 2013, LOFAR: The LOw-Frequency ARray. *A&A*, 556, A2.

Wilkinson, M. D., et al., B. 2016, The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

https://doi.org/10.1017/S1743921325104808 Published online by Cambridge University Press