# SDS@hd – Scientific Data Storage

Martin Baumann, Oliver Mattes, Sabine Richling, Sven Siebler and Alexander Balz

Computing Centre (URZ), Heidelberg University, Germany

SDS@hd is a central storage service for hot large-scale scientific data that can be used by researchers from all universities in Baden-Württemberg. It offers fast and secure file system storage capabilities to individuals or groups, e.g. in the context of cooperative projects. Fast access is possible from data generating facilities like microscopes as well as from data analysis systems like HPC systems. Data protection requirements can be fulfilled by data encryption and secure data transfer protocols. The service is operated by the Heidelberg University Computing Centre.

## 1. Introduction

SDS@hd [1] is a service that provides large-scale storage for scientific data and is meant to be used for "hot data" – data that is frequently accessed and worked with. The service can be used by researchers at most public higher education institutions in Baden-Württemberg. User authentication and authorization are implemented in terms of the federated identity management in Baden-Württemberg bwIDM [2] allowing researchers to use the existing ID of their home institution transparently for this service. The SDS@hd service website [1] provides further information about the technical and institutional requirements and the registration process.

## 2. Features of SDS@hd

**Tailored to the safety of research data**    Research data is a precious resource and should be stored using a trustworthy service to keep it safe from prying eyes. Data handled via SDS@hd is stored in an appropriate environment at the Heidelberg University Computing Centre (URZ). It is protected by state-of-the-art technologies, encryption as well as restrictive access and data policies.

**Ideal for collaborations**    SDS@hd is useful for researchers from different departments or institutions who want to work together. They can join a collaborative storage project and store their research data at a single spot. Using a web interface, the storage project owner can manage user groups and user roles and can thus determine who is allowed to access which parts of the data storage.

---

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00026851` veröffentlicht.
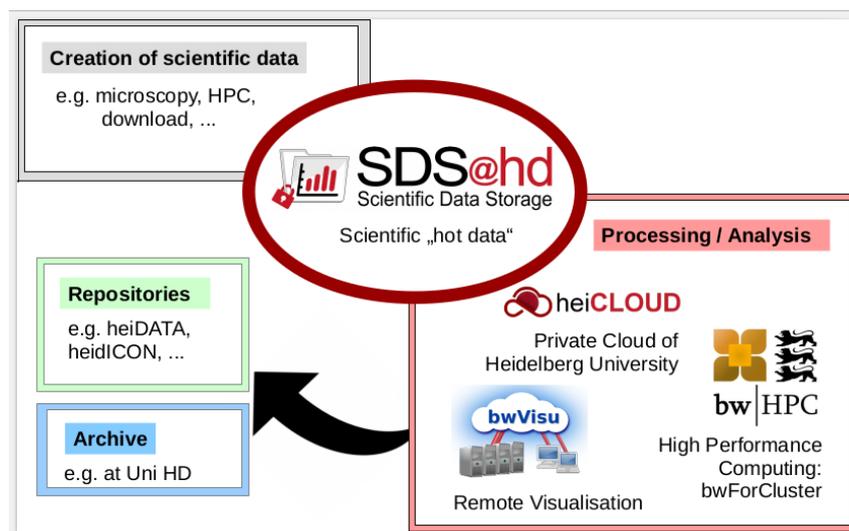
Figure 1.: SDS@hd supports the entire scientific data workflow.

**Storing HPC and visualisation data**   The SDS@hd service is connected to a range of data-intensive computing resources also available through the Heidelberg University Computing Centre. These resources – such as the bwForCluster MLS&WISO or the bwVisu visualisation service – can automatically obtain or save data to and from the SDS@hd storage infrastructure. General access to SDS@hd, e.g. from the bwUniCluster and other bwForClusters or from your notebook, is possible via the protocols SMB (2.x/3.x), NFSv4 (Kerberos) or SFTP. Users no longer need to find their own large-scale storage solutions.

**Support for the whole data life cycle**   SDS@hd aims to serve the entire scientific data workflow (see Fig. 1). When research projects are finished, support is provided for the transfer of data to general or community-specific archives and repositories.

## 3. Hardware and Software

The data is stored on the **Large Scale Data Facility (LSDF2)**, a state-of-the-art storage system located at the URZ's main server rooms in Heidelberg. The system is associated to a project between URZ and the Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT). The project aims to provide modern and valuable storage services for researchers.

The system is gradually extended and will provide up to 25 petabytes of storage space in 2020. A range of the newest HDDs interspersed with several solid-state drives (SSDs) guarantees high access speeds. Even large numbers of smaller files – usually a time-consuming challenge for storage systems – can be swiftly processed.

LSDF2 is made for high availability: An efficient RAID concept prevents any data loss, and a backup power system weathers any potential power outages. There is no single point of failure in the system. Additionally a backup concept allows to access the data, even in a disaster situation without any time consuming data restore needed.

On LSDF2, the high-performance clustered file system software "Spectrum Scale" (formally known as "GPFS") developed by IBM is used. It provides concurrent and fast file access to applications via various protocols including NFS and SMB. It offeres several features like high availability, disaster recovery and shared access to file systems from remote Spectrum Scale clusters. Additionally, tools for management and administration of storage clusters are contained. In total, Spectrum Scale is a solid and flexible basis for high-quality storage services for research.

## Acknowledgements

## Bibliography

[1] SDS@hd – Scientific Data Storage. `https://sds-hd.urz.uni-heidelberg.de`

[2] Föderiertes Identitätsmanagement der baden-württembergischen Hochschulen. `http://bwidm.de`

[3] Hartenstein, H., T. Walter, and P. Castellaz. "Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste." Praxis der Informationsverarbeitung und Kommunikation, Band 36, Heft 2 (2013): 99-108.