Managing large research data with SDS@hd

Sabine Richling, Sven Siebler, Alexander Balz, Robert Kühl and Martin Baumann

University Computing Centre, Heidelberg University

The scientific data storage 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 for individuals and groups. The service is operated by the Heidelberg University Computing Centre and running in production since 2017 with a continuously growing number of users and storage projects. Access management can be done via a predefined set of roles and also based on access control lists on the filesystem level enabling researchers to share data in a collaborative fashion.

1 Introduction

In many fields of research, the capacity of generated scientific data is enormous and continuously growing. This is a consequence of technical progress in data generating devices (e.g. high-throughput microscopes, telescopes and genome sequencers) and increasing performance capability of computer systems (e.g. high-performance compute clusters or cloud systems). Research projects have additional requirements including group and access management for co-operational setups, data protection for projects dealing with sensitive data, the demand in flexibly sharing data with others, data publication and long-term preservation.

The scientific data storage service SDS@hd [1, 2] builds on top of the second generation hardware of the "Large Scale Data Facility" (LSDF2) offering a fast and large-capacity storage backend for such needs. The LSDF2 is part of the state of Baden-Württemberg's concept for data-intensive services bwDATA [3].

SDS@hd has been developed to improve the value of available data storage resources in the context of research projects. It offers processes for user registration, role management, access management and collection of information for reporting. The service is open to all scientists at Baden-Württemberg's universities in the sense of a "Landesdienst".

SDS@hd is tailored primarily to those phases in the research data life cycle in which frequent and fast data accesses are necessary. In this phase, users can profit from the fast direct connection to the local high-performance compute cluster "bwForCluster MLS&WISO" [4] and can share their data with other registered users of SDS@hd.

For data publication or long-term preservation of research data, appropriate platforms must be used in addition to SDS@hd, see Fig. 1. Several community-specific platforms

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13759 (CC BY-SA 4.0)

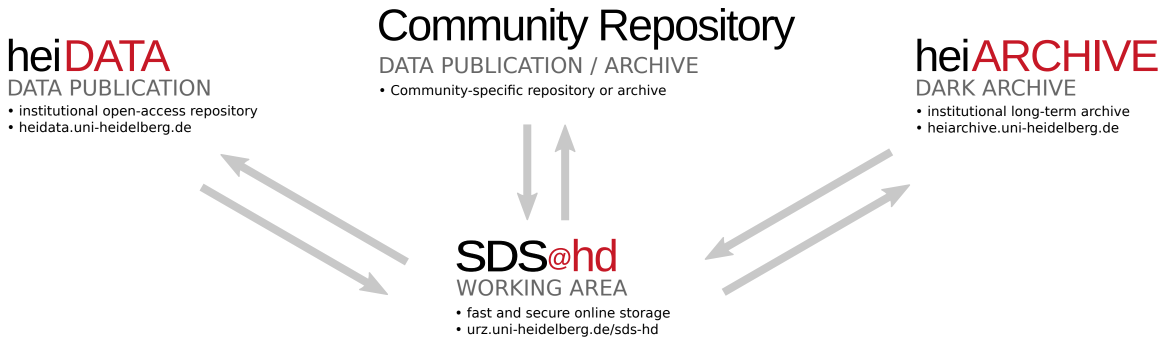
Community Repository
DATA PUBLICATION / ARCHIVE
• Community-specific repository or archive
heiDATA
DATA PUBLICATION
• institutional open-access repository
• heidata.uni-heidelberg.de
heiARCHIVE
DARK ARCHIVE
• institutional long-term archive
• heiarchive.uni-heidelberg.de
SDS@hd
WORKING AREA
• fast and secure online storage
• urz.uni-heidelberg.de/sds-hd

Figure 1: For data publication or archiving, data has to be transferred from SDS@hd to dedicated services. The institutional services heiDATA and heiARCHIVE of Heidelberg University facilitate open-access data publication and long-term preservation and are connected to SDS@hd.

exist for many disciplines and are often preferred over alternatives (e.g. institutional platforms). This is the case since appropriate features for the respective community are available including support for specific metadata standards which simplifies the locating and reuse of data. There are needs where community specific platforms cannot or should not be used. For such situations, many universities and other research institutions are operating institutional repositories for their members. In this context, Heidelberg University offers the institutional open-access data publication service heiDATA (`https://heidata.uni-heidelberg.de`) and the upcoming institutional long-time preservation service heiARCHIVE [5]. Both services are connected to SDS@hd and allow for metadata management and allocation of persistent identifyers. Whenever data is transmitted to and registered in such platforms, more or less requirements related to the data structure, file formats, and standardized metadata have to be fulfilled which most often requires some manual steps during the ingest process. For the transfer of the data from or to such services, different access protocols are available in SDS@hd and additional technologies will be added as required, cf. [6].

Subsequently in this publication, we outline the access management followed by some statistics of users and projects of SDS@hd. We briefly outline how the service is used in different research projects based on a selection of use-cases. We present a statistic of publications by users of SDS@hd and, finally, describe some planned developments.

2 Access management

For the management of the storage resources and accesses, storage projects are introduced as the organizational units of SDS@hd, denoted “Speichervorhaben” or short “SV”. One such storage project consists of one storage share with a defined quota and corresponds to a dedicated group of users. Members of the same storage project are able to share data easily while the access of non-members by default is not possible. There is a responsible person for each storage project who takes responsibility for the stored data and also for the reporting (e.g. for evaluations by the DFG). The responsible person can manage members and roles for his or her storage project via a web management tool. A fine-grained access management within one SV on the level of user roles is available. For example, a guest role allows only restricted access to the project, which can be used to exchange data with external collaborators in a dedicated download/upload area of the project. Additionally, it is possible to use ACLs on filesystem level to customize access permissions for members of a storage project.

The use of SDS@hd is possible in general for all bwIDM member organizations and involves a registration procedure on user level. bwIDM [7] is the federated identity management of Baden-Württemberg's universities which is realized as sub-federation of the DFN-AAI [8]. This technology allows researchers to use the ID of their home institution when using SDS@hd. The access is controlled by the entitlements "sds-hd-user" and "sds-hd-sv" which are granted by each organization for their own staff and students. A concept for the creation and management of guest accounts for research partners outside of bwIDM is in development.

3 Usage statistics

In the last 5 years the number of storage projects and users of SDS@hd was continuously growing. In 2017, the service started after a migration from a previous storage service with little over 50 projects and 150 users. Meanwhile, there are 850 registered users which are organized in 235 storage projects (see Figs. 2,3).

Looking at these numbers, it has to be noted that the number of persons who profit from (and de facto use) SDS@hd is higher. This is due to the fact, that several groups and projects implement a "proxy concept" where some (potentially high) number of end-users access the data via a proxy service, e.g. an analysis platform. The connection between the proxy service and SDS@hd is realized on behalf of only one registered user and is therefore only counted as one user in the statistics. The actual number of such end-users is not known but can be substantial. Such proxy services are implemented e.g. by core-facilities providing storage space to their customers for microscopy or sequencing tasks, or web-services like Omero (`https://www.openmicroscopy.org/omero/`).

As a central storage service for research projects, SDS@hd is used by researchers from different scientific fields. The user communities of SDS@hd and their shares in storage projects are shown in Fig. 4. Life sciences, medical sciences, and digital humanities have the largest shares. Together they have a share of 81%. The remaining share covers scientific computing, astrophysics and further disciplines.

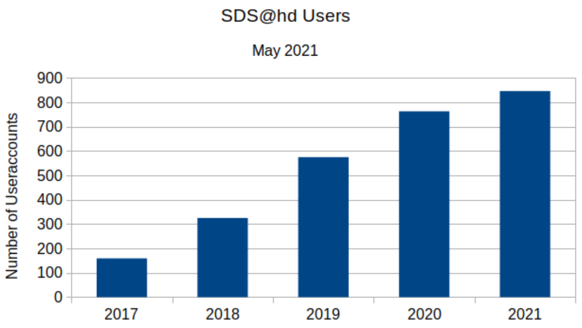

Figure 2: Number of SDS@hd users.

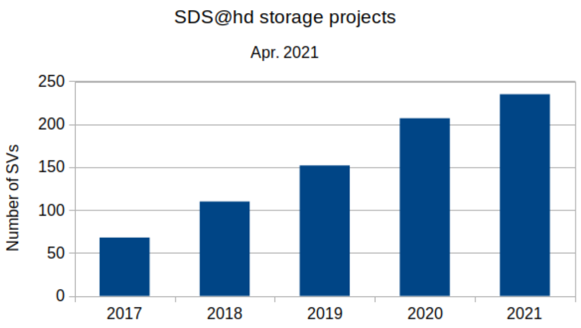

Figure 3: Number of SDS@hd storage projects.

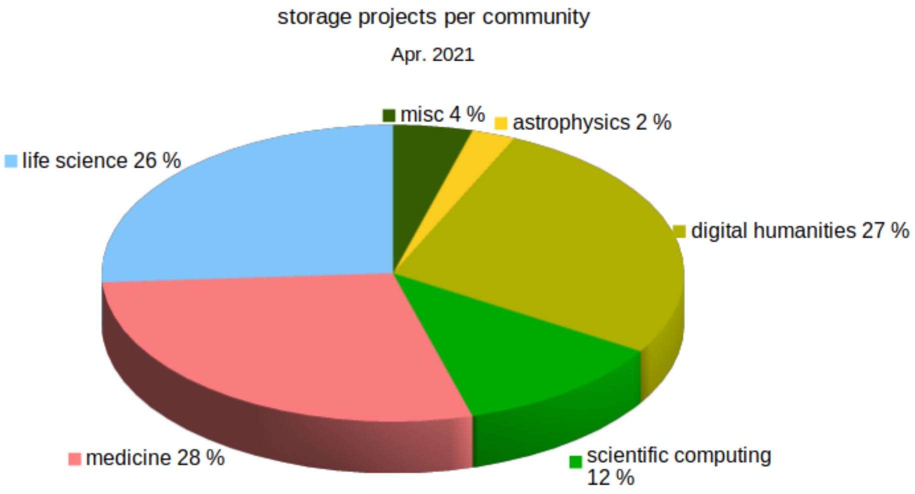

Figure 4: SDS@hd storage projects by communities.

4 Use-Cases

The scientific communities use SDS@hd for different workflows. Here we sketch a few exemplary use-cases for data-intensive research processes.

In life sciences and medical projects different types of microscopes and sequencers create data with high throughput. For example in the field of cryo electron microscopy new generation instruments are capable of producing several TB of data per day. The data are first cached locally and then transferred to SDS@hd. The original images must be available for several months for 2D and 3D structure analysis [9]. Medical projects with similar workflows deal with 3D images from expansion microscopy to identify structures in organs [10] or with high-resolution images of neuronal tissue samples for the analysis of neuronal structures and functions [11]. Other workflows in medical sciences use SDS@hd as download area and workspace for large scale genetic data from external sources.

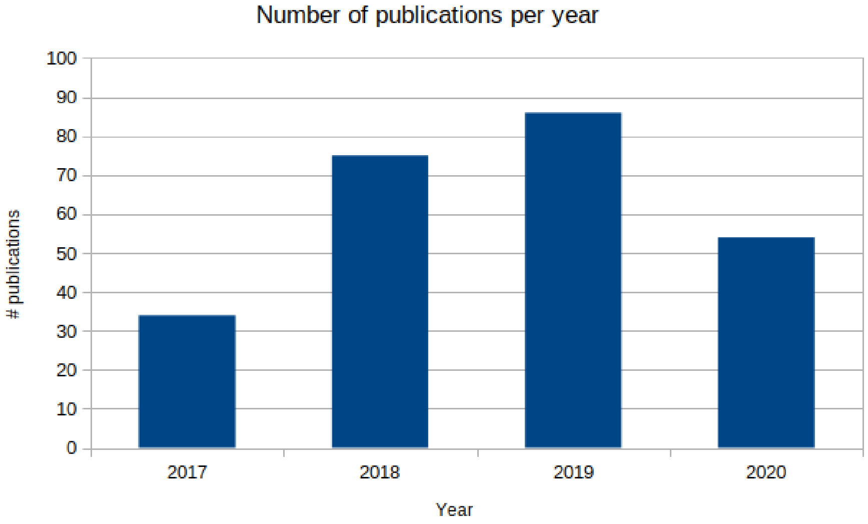

Figure 5: Number of reported publications 2017 - 2020.

In the humanities the digitalisation in historical sciences, archaeology, and linguistics creates big data and new challenges for data management. In this contect, a central storage solution as SDS@hd is essential for collaborative projects. For example, during an archaeological excavation digital and analog data are collected. Analog data like handwritten notes require digitalisation. All data are finally transferred to a storage project in SDS@hd for joint analysis.

In the field of scientific computing large scale data are produced in numerical simulations. Often it is required that the spacial distribution of physical quantities must be saved at many points in time to analyse the dynamic evolution of processes, for example in astrophysics [12].

For many use-cases the direct access to SDS@hd from the bwForCluster MLS&WISO becomes increasingly important for the production and post-processing of large simulation results as well as the analysis of images or genetic data. This development is reflected in a growing usage of SDS@hd in compute jobs.

5 References related to SDS@hd

Overall, a total of 249 publications related to SDS@hd were reported by users. Figure 5 shows the number of publications over the time. The rapid increase of publications shows a quick adoption of SDS@hd by its users and indicates the great value of this service for research projects. The amount of publications reported for 2020 is comparatively low, however it is expected to increase. The reason for this is that most publications are typically reported during the application of the storage project's extension which needs to be done once per year.

6 Future Work

In the future, the interplay of SDS@hd and services for publication repositories and long-term archives will be further developed and is planned to be the topic of a subsequent publication. In the context of the project bwHPC-S5, a data federation between the heterogeneous and distributed systems and services within Baden-Württemberg is in development [3]. SDS@hd will be under further development and is planned to be integrated into this emerging federation which promises simplified technical and organizational transitions from one IT service to another and by this increases its value for research.

Acknowledgements

The authors gratefully acknowledge the large scale data facility (LSDF2) supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. Federated user support for SDS@hd is funded via bwHPC-S5 by MWK.

Bibliography

[1] Baumann, M., V. Heuveline, O. Mattes, S. Richling, S. Siebler. “SDS@hd – Scientific Data Storage.” Proceedings of the bwHPC Symposium 2017, Universität Tübingen (2018).

[2] Baumann, M., O. Mattes, S. Richling, S. Siebler, A. Balz. “SDS@hd – Scientific Data Storage.” Science-Tage 2019 – Data to Knowledge, Heidelberg: heiBOOKS (2020). https://doi.org/10.11588/heibooks.598.c8444

[3] Hartenstein, H., T. Walter, and P. Castellaz. Schneider, G., V. Heuveline, K.H. Horstmann, B. Neumair, T. Hätscher, A. Pfister, J. Beutner, M. Resch, T. Walter, S. Wesner, R. Dorn, M. Holst, M. Steuert, and J. Last. “Rahmenkonzept der Hochschulen des Landes Baden-Württemberg für datenintensive Dienste – bwDATA Phase III (2020-2024).” Publikationssystem Universitätsbibliothek Tübingen. http://dx.doi.org/10.15496/publikation-55923

[4] Richling, S., M. Baumann, S. Friedel, and H. Kredel. ”bwForCluster MLS&WISO.” Proceedings of the 3rd bwHPC-Symposium: Heidelberg 2016, pp. 103-107 (2017). https://doi.org/10.11588/heibooks.308.418

[5] Baumann, M., F. Heß, L. Maylein, T. Mechler, B. Scherbaum, and E. Volkmann. “heiARCHIVE, a long-term preservation service at Heidelberg University.” E-Science-Tage 2021 – Share Your Research Data, Heidelberg: heiBOOKS (2021).

[6] Baumann, M., F. Bösert, S. Siebler, P. Skopnik, and J. E. Sundermann. “Entwurf einer Infrastruktur für den Datenaustausch großer Forschungsdatenmengen mittels WebDAV, FTS3 und OIDC.” E-Science-Tage 2021 – Share Your Research Data, Heidelberg: heiBOOKS (2021).

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

[8] DFN-AAI - Authentication and authorization infrastructure. `https://www.aai.dfn.de/en/`.

[9] Zupa, E., A. Zheng, A. Neuner, M. Würtz, P. Liu, A. Böhler, E. Schiebel, and S. Pfeffer. “The cryo-EM structure of a γ-TuSC elucidates architecture and regulation of minimal microtubule nucleation systems.” Nature Communications 11, 5705 (2020). `https://doi.org/10.1038/s41467-020-19456-8`

[10] Cinzia, B., A. u. M. Khan, T. Picascia, Q. Sun, V. Heuveline, and N. Gretz. “New technical approaches for 3D morphological imaging and quantification of measurements.” The Anatomical Record 3, 10 (2020). `https://doi.org/10.1002/ar.24463`

[11] Klevanski, M., F. Herrmannsdoerfer, S. Sass, V. Venkataramani, M. Heilemann, and T. Kuner. “Automated highly multiplexed super-resolution imaging of protein nano-architecture in cells and tissues.” Nature Communications 11, 1552 (2020). `https://doi.org/10.1038/s41467-020-15362-1`

[12] Pellegrini, E. W., S. Reissl, D. Rahner, R. S. Klessen, S. C. O. Glove, R. Pakmor, R. Herrera-Camus, and R. J. J. Grand. “Warpfield population synthesis: the physics of (extra-) Galactic star formation and feedback-driven cloud structure and emission from sub-to-kpc scales.” Monthly Notices of the Royal Astronomical Society 498, 3 (2020). `https://doi.org/10.1093/mnras/staa2555`