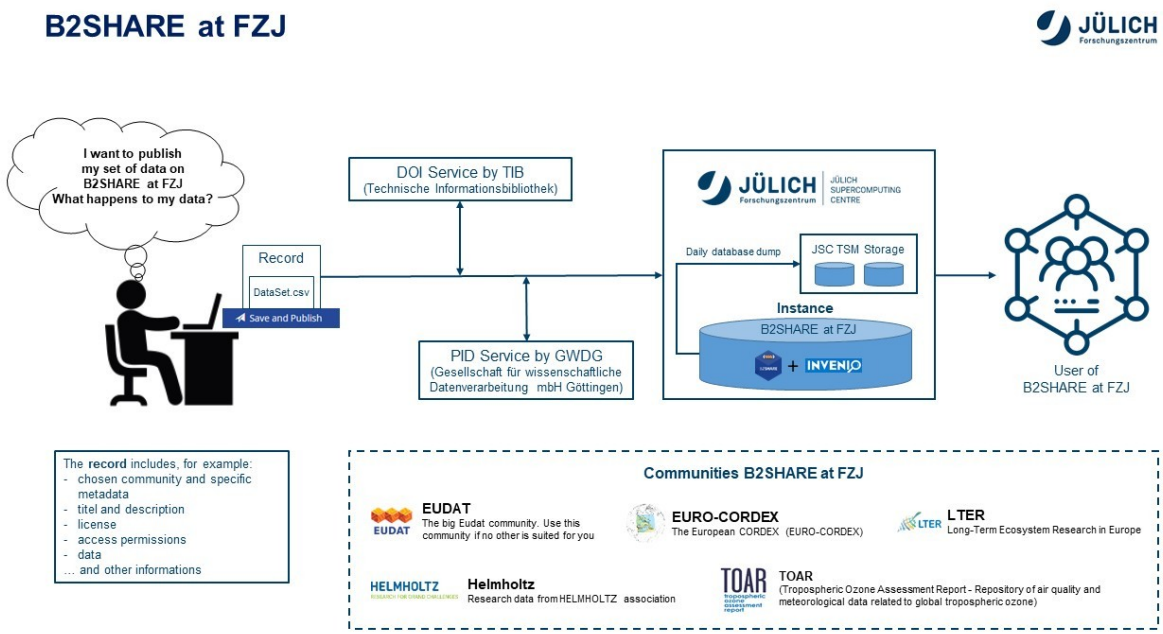
Data Management Plan for B2SHARE at FZJ

Autor: Sander Apweiler, Björn Hagemeier, Eleonora Epp

Date: 02.12.2024

1 Introduction

The B2SHARE at FZJ (Forschungszentrum Jülich GmbH) is offered by the JSC (Jülich Supercomputing Centre) and provides researchers a solution for storing, sharing and publishing research data from diverse research areas. Data is published in a record that is registered with a PID and a DOI. The record can handle research data in a single file or split data across multiple files. Each file is registered with a PID. The additional metadata of the research data is stored in the record as well.



As a generic data repository, we do borrow concepts from the OAIS reference model. Data ingest is under the control of the user, who adds descriptive information in the form of metadata to their so-called deposits. B2SHARE as a software ensures secure storage of the data by creating checksums at ingestion and regularly verifying correctness of all data according to these checksums. Furthermore, the repository's data and metadata are both backed up in order to allow for recovering data in various situations, be they a full crash of the entire system or individual data corruption due to storage media failure. Access to the data is possible through downloading deposits via a web interface. Ultimately, all data are intended to be publicly available to everyone. However, user-defined embargo periods can limit this availability temporarily. All data of the repository can be searched, including the option to search for data from overarching search engines such as B2FIND.
Long-term preservation of the data is ensured by migrating the data as-is to any new storage infrastructure over time. This ensures bit-level preservation of all data, but does explicitly not include preservation at the application level, which we consider to be a responsibility of users and their communities.

2 Infrastructure, Storage and technical backup

The B2SHARE software stack is deployed using docker containers and managed via docker compose. In addition to B2SHARE itself, the software stack contains Elasticsearch for indexing, PostgreSQL for storing record metadata and user accounts, and nginx as a webserver for serving the web UI. B2SHARE is operated on a virtual machine within an OpenStack cluster, using the JSC's central storage infrastructure as a back-end. IBM Storage Protect[1] is used as the backup system for all data. Beside the instance for publishing research data, a dedicated instance for testing purpose is provided and linked from the production instance. B2SHARE is a publication service; no security is guaranteed for sensitive data.

3 Documentation and data quality

The documentation standard has been well described for B2SHARE users in general by B2SHARE EUDAT on the documentation website[2].

Each community can create their own metadata schemas comprising mandatory and optional fields to be filled before publication.

In terms of data quality, B2SHARE ensures that the data remains unchanged through check summing. We also ensure that the data that has been uploaded is available for a long time (long-term archiving).

The content level is not covered by B2SHARE. The owners are responsible for the correct content and the usability of the data, data sets, etc. The community can send a report for each publication if it violates rules or is not usable. The owner can correct the data records at any time with a new version number. The owner and / or the community are responsible for data curation.

4 Setup and Operation

The instances are managed using the configuration management system Puppet[3] which applies well-defined configuration for the services and checks for changed configuration regularly. This allows to apply the configuration updates of the underlying server also from administrators who are not familiar with the B2SHARE software itself. It also allows an easy and fast deployment of a server replacement in case of problems with the old server. The server is monitored using Icinga[4] monitoring system, which sends notifications to the administrators upon service status changes. A daily backup of the research data, metadata and database export is performed.

[1] https://www.ibm.com/products/storage-protect

[2] https://doc.eudat.eu/b2share/forusers/

[3] https://www.puppet.com/

[4] https://icinga.com/

5 Legal obligations and framework conditions

Legally, we rely on EUDAT B2SHARE, from whom we obtain the software. Use Policy and conditions of use: https://www.eudat.eu/eudat-cdi-aup .

B2SHARE is a publication service, there are no restrictions on access with the exception of private datasets, for which only metadata can be viewed. If there is an embargo period, the data will not be published automatically after the deadline expires, but the owner must take action. In cases of copyright or other infringements, data can be removed from B2SHARE.

It is in the owner's responsibilities to take care of the data. In the case of ownership extensions, ownership cannot be transferred to other owners. This function is planned for the future.

6 Risk Assessment

B2SHARE at FZJ is a general data publication service at Forschungszentrum Jülich. Data, metadata, software and publications are published here and made available to other users via https://b2share.fz-juelich.de/. All files are stored at Jülich Supercomputing Centre for long term archival, including backup to keep any risks as low as possible. For this we have carried out an internal risk assessment. The data risks are evaluated on the basis of the risk assessment.

We take action against each individual risk factor and carry out a new risk assessment every two years. The last assessment has taken place in 2024, the next assessment will be due in 2026.

The risks are evaluated on the basis of the risk assessment matrix from Matthew S. Mayernik: https://datascience.codata.org/articles/10.5334/dsj-2020-010.

In the following, the main risk categories for the B2SHARE at FZJ services are given with the relevant risk factors including the estimated degree of control, the estimated impact on users, and the countermeasures taken proactively:

6.1 Data Risks

In the following sections, we provide a risk severity and and impact for each of the risks. However, because B2SHARE at FZJ is a generic repository, we identified these attributes only for a subset of risks, as they may be influenced by the actual data that is stored by the users.

Lack of metadata & documentation (risk: high, impact: low)

Publishers of data on the B2SHARE at FZJ service are responsible for the accuracy and usability of their data. The B2SHARE at FZJ and the EUDAT B2SHARE documentation is carefully maintained.

All uploads must be accompanied by metadata following a schema. A generic schema is provided by B2SHARE and its mandatory fields must always be filled. B2SHARE supports additional schemas for each community, potentially adding further mandatory fields. It is up to the communities to define schemas including optional and mandatory fields. Filling the fields upon uploading data is the responsibility of the user doing the upload. Data sets can only be published once the metadata has been filled completely, i.e. including all mandatory fields.

Whereas metadata can be updated at a later point in time and does not lead to a new version of the corresponding record, changes in the actual data will always lead to a new version of a record. On the other hand, users have the choice of creating new versions for metadata updates. The decision whether or not to create a new version is the user's responsibility.

Data mislabelling (risk: low, impact: high)

To avoid data misidentification we accept entries through the B2SHARE form to upload data, the B2SHARE form receives entries related to the publication. The unique DOI and PID, generated by B2SHARE at FZJ, make the data clearly identifiable.

Accidental deletion (risk: low, impact: medium)

Only very few skilled people have elevated privileges on the B2SHARE at FZJ service. System updates and major database updates are planned by at least two people. The infrastructure is set-up so that it rarely requires manual intervention. Restoring data from backup copies is possible and thoroughly documented.

Poor data governance

New staff members receive intensive training, and a strict control of access rights and responsibilities is enforced, allowing only experienced personnel have access to the data. The potential deterioration of data governance is minimized through regular retraining or replacement of personnel.

Lack of planning

To avoid overloading measures for the transfer of responsibilities are established. Written documentation and planning are important to us, and regular team meetings are held to discuss issues, progress, and plans.

File format obsolescence

Regular user forums and user interaction will allow to foresee requests and code development can be planned early and will prevent file formats from becoming obsolete. Metadata are not stored as files but in a database, i.e. a conversion tool can easily be made available. Durable file formats have been chosen for the database. The versioning option in B2SHARE allows creators to upload new formats if the old format became obsolete.

Over-abundance

Over-abundance is the risk of storing too much data without filtering, such that the value or quality of the data is reduced as a consequence. It is up to the users of B2SHARE to avoid this risk, the repository itself does not mandate any specific procedures in this regard.

Furthermore, if users start storing excessive amounts of data in the repository, administrators would contact them to take measures to avoid the depletion of available storage.

6.2 Physical Risks

Media deterioration

To prevent the primary risks associated with the ageing of hard disks or tapes, a comprehensive backup strategy is essential. All file systems, including those for the operational database, are stored

on IBM Spectrum Scale Storage disks. Hardware is closely monitored, and any signs of degradation prompt immediate replacement. In the event that such failures result in data loss, the affected data can be restored from backup.

Storage hardware breakdown

To counteract this risk, the following measures are implemented: continuous monitoring of hardware components, a hot-swap strategy, regular acquisition of new hardware, and regular updates of databases and services.

Cybersecurity breach

Elevated security measures are implemented at JSC and regularly reviewed and updated. As long-term HPC service provider there is strong awareness of cybersecurity issues. In case of a breach the B2SHARE at FZJ services are cordoned off. The instance and database of B2SHARE at FZJ can be reinstalled from a trusted backup version in case of malicious attacks.

Bit rot and data corruption

Our database technology includes backup measures for the B2SHARE at FZJ database. Depending on the severity of the problem, data is supplemented by the backup.

In order to avoid silent data corruption, B2SHARE creates checksums of uploaded data and checks data for consistency at regular intervals.

Malicious attacks

Access to the hardware hosting the B2SHARE at FZJ infrastructure is regulated and only possible for designated personnel. Security measures include locked doors with an entry system logging who entered the machine hall. Furthermore, the hardware is placed on a secured campus with manned gates.

Human error

The hardware is operated by trained personnel with long experience in operating complex high performance computer architectures.

Catastrophes

For the machine halls early warning systems are in place as well as a fire-extinguishing system (Argon) in the main one. A fully equipped fire brigade specialising in all types of fires is located at FZJ. The fire department is located about 650m away from the supercomputing centre.

6.3 Human Risks

Lack of use

Regular user consultation to inform potential users and their communities. We are working regularly to improve the database and the documentation for our existing users.

Loss of knowledge around context or access

Continuous training of personnel, building redundancy in terms of knowledge to operate the B2SHARE at FZJ instance are in place as well as building up knowledge about the B2SHARE at FZJ

infrastructure in our Team. Software and data is archived or published so that the system can be rebuilt by trained personnel.

6.4 Organisational Risks

Loss of funding for storage

Funding to replace the storage currently underlying the Cloud-based hosting environment has recently been secured and lasts until at least 2030. Should funding be lost in the future, the instance could be set up and used at another Center of the Helmholtz Association of German Research Centres[5], who have built up an IT infrastructure distributed across centres in a collaborative effort[6].

Legal status for ownership and use

We accept all data that can be used within the community. Users publish their data so that it can be used. The legal status, such as licenses, must be indicated with each publication.

Political interference

Recent years have shown that even stable democracies can be at risk, which can lead to radical changes of politics and policy making, thus also endangering the freedom of science. One infamous example of this happening was the Trump administration's interference with and attempts to influence scientific results in their favour[7]. We do not currently foresee any eviction schemes in such cases. Within Europe, the participation in EOSC may provide such opportunities, but they are not yet tangible and highly dependent on the actual future political threat.

[5] https://www.helmholtz.de/

[6] https://www.hifis.net/

[7] https://en.wikipedia.org/wiki/Trump_administration_political_interference_with_science_agencies