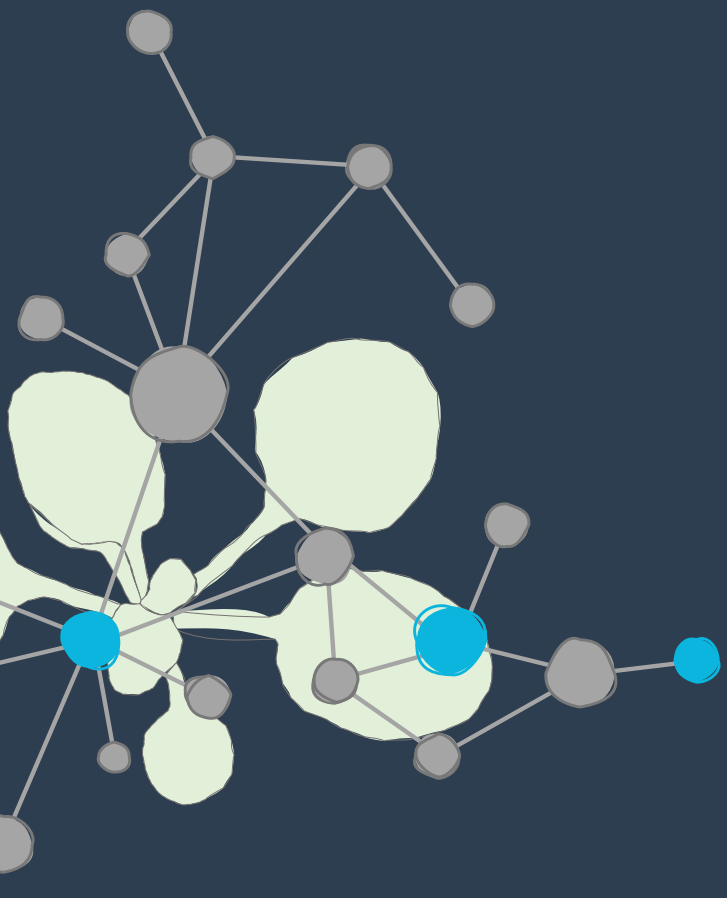
Data Sharing

Dominik Brilhaus



DataPLANT

The merits of data sharing

Research is a collaborative endeavour that builds on the interaction and efficient knowledge exchange between different researchers. We share research data to get input from peers and elaborate, initiate or expand putative or existing collaborations. Data sharing allows us to save time and resources, e.g., by finding partners to plan or perform investigations together, sharing common pipelines for data analysis or prevent redundant or overlapping investigations, simply by knowing what other peers might already investigate. Sharing research data is thus the key to every successful research project.

However, sharing data is frequently hindered even between researchers of close surroundings. There may be legal reasons, including unclear policies from funding agencies or institutions: *"Who am I allowed to share my data with?", "How do I handle data requiring specific precautions for data security or intellectual property rights?"*. Social or emotional reasons might occur, if researchers might not know about peers interested in their own data: *"How do I know, who would like to see my data, if they do not know it exists?"* or are afraid to "lose" their data: *"Once I share my data, someone else will publish and get credit for it"*. Recent developments of open science have boosted scientific advancements. However, it is a common misconception of the FAIR principles of data stewardship that accessible data equals public and openly accessible data.

Most researchers however want to share their data and are very aware what data to share with whom, but face technical or even financial issues: *"Where and how can I securely share and integrate research data of multiple types, originating from multiple sources?".* The sheer amounts of data and data types produced during complex multi-party investigations can easily become overwhelming to handle, costly to store, or limited by storage capacities, especially when proper data protection mechanisms are employed.

The one-stop-shop does not exist

Today many options for sharing and collaborating on data are available and often consciously or incidentally integrated into daily research routines. These include prominent open source or commercial cloud platforms like nextcloud, google drive, dropbox, onedrive and many more. While these are great for synchronous collaboration on typical office data, text files, presentations or simple calculations, they offer limited capacities for data analyses, especially those required for large-scale or complex scientific data. Other solutions specifically designed to accommodate scientific data include electronic lab notebooks to document daily lab routines or platforms like galaxy and omero to analyze and share data from omics or imaging experiments, respectively.

To varying extents, these platforms offer a mix of options for local and remote, asynchronous and synchronous collaboration, often supported by automated version-control to track file version history.

Different modes and control of access to the data and different solutions for storage sites exist to suit various aspects of data security and property rights.

For research individuals or groups the data sharing dilemma often lies in the fragmentation of data shares. The more projects collaborating on data of different domains, types or formats and the more people and groups involved, the more platforms are being used, resulting in a fragmented and barely accessible or efficiently manageable data landscape. As a consequence, (un)published data is still mostly shared through conventional routes, such as direct communication between peers via email, instant messaging, virtual and live, personal or group, meetings or presented in more formal contexts such as reports and symposia. As these formats frequently focus on late-stage or final research outputs only, they diminish the chances for collaborations early in an investigation.

Changing the dogma from tool-bound to data-centric: Good data sharing

Trying to find the tool or platform most suitable for the project or data to be shared always depends on the context and is innately erroneous, leading to increased fragmentation. Now, how can the data fragmentation be resolved without siloing everything in one place, i.e. yet another platform? In order to set loose from platform dependency, one could flip the data sharing habits inside-out and switch from the tool perspective towards a data-centric perspective. Instead of trying to enforce the use of a specific platform for data sharing, one could use a data format suitable to and migratable between a wide range of tools and purposes.

In order to support FAIR data sharing, such a data format requires high flexibility to be adoptable to many data types and sources, long-term persistency through independence of (i.e. extension or conversion to) specific data formats and scalability to increasing data amounts. Federated data storage and access allows secure, trusted data sharing with involved parties from different locations across institute borders. Data protection is further granted through geo-redundant backup mechanisms. In combination with a version-control system to follow file change history, the federated authentication and authorization system allows to control data access and contribution for proper crediting and provenance tracking. Data sharing is enabled throughout project lifetime – from idea to unpublished data to publication –, by structuring the data in a defined format packaged with descriptive metadata and licenses to provide technically and legally clear terms of data (re-)use. From there, data publication comes with as little effort as assigning a persistent identifier without any need to adapt the data once the associated manuscript is published.

How does DataPLANT support me in data sharing?

The following table gives an overview about DataPLANT tools and services related to sharing data. Follow the link in the first column for details.

Name	Type	Tasks on data sharing
ARC (Annotated Research Context)	Standard	**Structure:**
DataHUB	Service	**Share:**

Register with DataPLANT

In order to use the DataHUB and other DataPLANT infrastructure and services, please sign up: with DataPLANT.