Storing Research Data

By Erik Axel Vollan, UiT The Arctic University of Norway

Date: October 2020

Manuscript:

In the life cycle of research data, you have finished the planning phase, and are entering the active phase. Now you will start collecting data and work on analysing them. Analysis may generate new data and may involve several rounds of collection. The data you collect need to be stored properly to minimise the risk of data loss, and to enable you to work on them.

The final stage in the life cycle is archiving. This is where you deposit your data in an archive and make them publicly available. Archiving and publication will not be covered here.

I would like to mention 3 aspects of data storage you should be aware of:

- **Confidentiality** of data means keeping data accessible only to those who are authorised to see them.
- **Accessibility** means that data are accessible to those who need them when they need them.
- **Integrity** of data ensures data are not changed or deleted without authorisation

I will return to these in the coming topics.

In order to store data correctly you need to know what kind of data you are working with. One important aspect is the sensitivity of the data. Sensitive data must be kept confidential. You will need to assess this while planning your project.

If your research includes personal data, you will be bound by privacy rules. For certain areas of research there will be laws and regulations that demand data be kept confidential. Examples can be research on health data, or research including minors. If your research data have commercial value, they may need to be kept confidential.

Most institutions have a formal system for data classification that will require you to classify data into one of several levels of sensitivity. The number of categories and naming of them will vary, so it is important that you seek information at your institution. As an example, UiT the Arctic University of Norway uses a system of 4 levels with increasing sensitivity: Open, Restricted, Confidential and Strictly Confidential. The levels are colour coded as shown.

Once you have classified your data, there will be implications for your handling of them. Your institution should provide specific rules for what classes of data are allowed on its different storage systems. This will of course be unique to the institution, and you need to check what applies at your institution.

Scientists work in a myriad of disciplines and use a wide variety of methods to collect their data. Data collection may involve recording an audio interview, using instruments in a lab or observing wildlife. The list is endless. It is important that data are recorded and stored securely during collection. If you work on campus, with access to computer networks and storage systems, you can record your data directly into university storage. If you are off campus, you may need to store data temporarily on a computer or some sort of storage medium like a USB drive. In this case, it is important to be aware that you now have data without backup, so make sure you keep that data safe until it can be uploaded to backed-up storage systems. Consider making a copy on a separate physical device if available to protect against malfunction of a device.

This brings us to another extremely important aspect of data storage, namely:

Safekeeping and backup

The data you collect during a project are the basis for your research. You and your employer put time, effort and money into collecting research data. Losing data will at best mean having to re-collect them, but some data may not be reproducible at all. Imagine for instance a biologist gathering data every summer for years and then losing the dataset for one or more years.

It is also easy to imagine the sinking feeling if someone were to steal your laptop containing your entire PhD, or an external hard drive malfunctions, losing all the data stored on it.

To avoid this, you should always strive to keep your data in a quality-assured storage system which is being backed up. Local IT at your institution will provide you with such systems. Use them! Usually you are granted a personal storage quota for your documents and files and specialised research data storage is often available too.

Using backed up storage is important to secure the accessibility and integrity of your research data. If you happen to delete a file in error, a backed-up system will allow you to retrieve the file. Having your data backed up also means you will be able to go back to earlier versions of files.

You may need to download data to your computer in order to work on them. If you do this, you need to establish a workflow uploading the results of your work to backed-up storage again. I know how easy it is to think “I’ll do it on Monday” or otherwise put it off, but you can quickly find yourself with a month of work stored without backup.

Most data can be stored in general storage systems. However, some research data has special needs and I will mention two specific cases here.

The first is research involving large amounts of data. In this case you risk exceeding the size quotas of your standard storage. Universities often have specialised research data storage systems with larger quotas for this purpose. If you find yourself needing more space, check with your local IT how to proceed. Be aware there may be procedures to go through for getting a quota.

You may also find services provided outside your home institution. Many countries have national e-infrastructures providing services like large-scale storage and supercomputing. To use such services, you may need to apply for quotas and this process can take time. You should identify such needs when making your data management plan and incorporate the timing into your project planning.

Sensitive data will need special vigilance in handling to avoid breaching confidentiality. A chain is only as strong as its weakest link, so it is important to focus on keeping confidentiality throughout the life cycle. This means keeping data secure all the way from collection, through analysis and finally archiving. Doing so is a combination of good procedures and using technical tools to protect data. It may be a good idea for you to write a little checklist describing in detail how you will handle data. If you collect sensitive data off-campus, you may need to store them on an encrypted hard drive or memory stick until you can upload them into a safe storage system. If you lose an encrypted device, the finder will not be able to read the data.

Your institution will have a policy regulating the handling of sensitive data and you need to check this. There will be guidelines for where and how to store such data, again you need to consult your institution.

If you collaborate on data during a project, it will be helpful to store them on a system that allows shared access.

If you and your collaborators work within the same institution, you will probably be able to share access to the same storage. Sharing access can be more of a problem if you collaborate across institutions and maybe also borders. You will need to check with your local IT department if there are systems that allow sharing with outsiders. It can be tempting to use cloud services to share data with collaborators. Since cloud services may store your data in data centres in other jurisdictions and even continents, you must check with your local IT what is allowed. If you are working with sensitive data, this is especially important.

Alternatively, the participants can work on their separate versions of a dataset. If you do work on separate copies of the same data, you must ensure you do not alter the data, ending up with differences between copies. Keeping a master version that you make working copies from can be smart.

If you end up having to send data to external collaborators, you must ensure that this transfer can be done safely and within regulations. Your local IT department will be able to help you here.