Many research projects and organisations have large volumes of research data that they need to store. As well as providing high performance computing systems for research, PDC also offers long-term storage for research data – currently this is primarily via our Mass Storage System (MSS) which is managed by IBM Spectrum Protect software (which was previously known as the Tivoli Storage Manager or TSM).

PDC is also heavily involved in an ongoing project to establish a storage system for Swedish research data using an approach based on the iRODS software for data management (for more information, see the iRODS section below). If you are involved in a research project that needs to store data long-term, you are welcome to contact PDC Support to discuss purchasing storage from PDC. As the PDC mass storage can be extended fairly easily and cheaply (by buying more tapes for the MSS and extra licenses for the software), this can be a more economical solution than other storage alternatives such as setting up a tape storage system dedicated just for a single project, or buying storage from commercial companies.

**Mass data storage at PDC**

PDC’s Mass Storage System is essentially a large library of magnetic tape cartridges that are accessed using a tape robot. This is a very efficient way to archive data that is not accessed frequently but that needs to be stored for a long time.

**Mass Storage System**

PDC’s MSS is essentially a large library of magnetic tape cartridges that are accessed using a tape robot. This is a very efficient way to archive data that is not accessed frequently but that needs to be stored for a long time.

**Some projects storing data at PDC**

The following projects have their primary data archives at PDC.

**CENTER-TBI** is a large European project “Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injuries” that aims to improve the care for patients with traumatic brain injury and identify the most effective clinical interventions for managing such injuries. The Odin satellite combines two scientific disciplines on a single spacecraft in studies of star formation and the early solar system (astronomy) and the mechanisms behind the depletion of the ozone layer in the Earth’s stratosphere and the effects of global warming (geoscience).

The Swedish Space Corporation, on behalf of the Swedish National Space Board and the space agencies of Canada (CSA), Finland (TEKES) and France (CNES), has developed the satellite for astronomers and atmospheric researchers in the participating countries.

**Prisma** is a Swedish-led satellite project that aims to develop and qualify new technology necessary for future science missions in space. Many of the future projects comprise formation flying and maintenance, so several spacecraft need to communicate and interact with each other with high precision. That requires exceptional accuracy in measuring and controlling the inter-satellite navigation.

**SNC-SENS** is a Swedish project that uses high performance computing resources for analyzing sensitive data. PDC is a partner in this project and provides a backup resource for the National Genomics Infrastructure (NGI), which includes backup of sensitive personal data.

The system is based on the IBM Spectrum Protect software and provides backup for the NGI facilities at the KTH Royal Institute of Technology and Uppsala University, and also acts as the backup of the NGI production systems which are operated by the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) at Uppsala University.

**The Human Proteome Resource (HPR)** is a large Swedish program that provides knowledge-based analysis and access to data on the functional aspects of the human tissue proteome. A basic description of a defined proteome is made available publicly and includes gene lists, analyses of expression patterns, and examples of protein expression on a cellular level. This data is archived and backed up at PDC (see also under IBM Spectrum Protect to the right).

**New iRODS storage underway**

In collaboration with other partners from the Swedish National Infrastructure for Computing (SNIC), PDC is developing a new service for research data storage. The new service will make it easier to manage research data and will be a significant step towards providing open access data. Researchers who make use of this storage service for live project data will be able to start generating metadata rights from the very beginning of their project. This will make it a lot easier to package data and archive it when their projects come to an end. The expectation is that archiving services will also benefit from these efforts immediately, as metadata is an important part of making data searchable, and is also useful when it comes to publishing data.

The Integrated Rule-Oriented Data System (iRODS) is open-source software that provides a comprehensive set of tools to support data management from the initial collection of data through to archiving and reusing the data. This is particularly important given the implications of the worldwide movement towards Open Science and Open Data Access. iRODS is supported and maintained by the Data Access and Discovery (D3) Consortium, and is used by research organisations and government agencies worldwide. Over the last few years the iRODS software has been subjected to very significant refactoring and reorganisation – the result of which is that iRODS is now being released as a production-level software distribution with commercial support, as well as a strong user community.

PDC’s iRODS storage service is currently available as a demonstration service, in collaboration with other HPC facilities across the globe. In the future, iRODS will be offered as a production service to any user organisation that wishes to adopt the iRODS system.

**Different projects and organizations are using PDC’s Backup and Archiving Solution which is connected to the mass storage system or MSS.** Data lands in the disk storage pools and, after a certain time, it is migrated to PDC’s MSS. Off-site backup of the system is performed twice a day to the National Supercomputer Centre (NSC) in Linköping. For example, the Swedish Human Proteome Resource (HPR) is using archiving at PDC and has stored about 370 TB of data which is mirrored off-site, in this particular case to the High Performance Computing Center North (HPC2N) in Umeå.