Data Archiving facilities and options

The Informatics Computing Team has been approached with a question of what facilities the School offers for archiving data. The simple answer to this is 'none' but that would make for a very brief article so let's take the time to define what we mean by archiving, explain why the School doesn't offer an archiving service and show what alternatives are available. As a first step, let's define the difference between an archive and a backup service (which the School does offer).

A backup is a copy of a given set of data taken with the aim of providing a means of recovering the original data if it is lost due to some malfunction or disaster. As time passes and the original data continues to change and be updated, the differences between it and the backup will grow, making the backup less and less useful as a means of recovering the data. At this point, a new backup needs to be taken. This is why the School backs up changed data every evening and deletes backups after 13 months to allow the media used for the backup to be recycled.

An archive, on the other hand, is a snapshot of a dataset at a particular moment in time, done with the view that having access to this data, as it was at this point in time, will be a useful thing to have at some point in the future. The exact definition of "some point in the future" is one of the many issues that makes archiving something that requires considerable thought.

As an aside, the School (and the departments which came together to form the School) used to pay lip service to archiving by permanently retaining a months worth of level 0 backups every year. As we will see, this procedure does not meet most, if not any of the requirements for a useful archive.

What is needed to make a data archive useful? First of all, we need a full description of what has been archived including what the data is, what format it is in, how it was created, how it can be used (i.e. what applications provide support for the format of the dataset), and to whom it belongs. We need to know where the data is located, and this may be a particular file server, a web site, a specialised repository or a magnetic tape stuck in the back of a dusty cupboard. Last, but very much not least, we need to be able to retrieve the data from its current location, a process which may range from finding out who knows the password required to access the repository the archive is stored within to locating a working tape drive compatible with the format of the (hopefully uncorrupted) tape.

This explains why our former policy of retaining a months worth of full backups every year was not a useful one. The information about the contents of each tape usually amounted to no more than a list of usernames or a list of partitions on a server which had long gone out of service. The problem of identifying the tape holding a given dataset can be understood. Even if the tape could be identified, if it was more than a few years old, there was every possibility that either the tape would have become corrupted or that the School would no longer possess a drive capable of reading the tape. This is why the decision was taken by the School to only retain backups for 13 months.

But proper archiving is important. Why? Well partially because it would be nice to be able to access all the important work done by the School's users over the years but mainly because archiving is something that grant funding bodies are becoming more and more concerned about. It is commonplace now for grant awards to specify for how long and in what matter, data created by a research grant remains available and accessible. Note an important point here. Archiving is not just about ensuring that data is available to access, it may also be about making sure that at some point in the future, the data is no longer available and possibly that it has been securely destroyed. Grant applications are expected to include a data management plan detailing how the researcher intends to meet the data access requirements of the grant funding body.

This is a lot for a harried academic struggling to meet a fast approaching grant application deadline to consider. Fortunately, help is at hand. Information Services offers a Research Data Service, "a suite of tools and support that helps staff and students be effective with their research data before, during and after their project". Their website http://www.ed.ac.uk/information-services/research-support/research-data-service should be the first place anyone wanting advice on how best to managed and protect their data should head for. In addition, The School's computing staff are always happy to offer more School-specific advice.