IT demo abstract for IDCC

Managing Research Data: Submitting BIG data to a DSpace repository

Ian Wellaway, Open Exeter Project, University of Exeter, UK

DSpace comes readily equipped with its own 'out of the box' submission tool which works well with small files and small numbers of files but how do researchers upload their precious large datasets?

The UK JISC funded Open Exeter project set out to understand how researchers at the University of Exeter manage their data.

As part of the project, researchers were surveyed about the amount of data they stored and how they stored it particularly once a project was finished. It was found that some research projects produced huge numbers of files with some massive file sizes and that these were often archived on local hard disks and external drives. To submit these datasets to our DSpace based institutional repository is not practical using the out of the box DSpace submission tool since it limits the user to one file at a time for upload over HTTP whilst the user waits. In addition transferring large files via such methods can be slow and prone to failure. DSpace does also support command line batch import of files providing they can be successfully transferred via some other means. SWORD provides a standardised way of interfacing with repositories including DSpace but also currently remains limited in its ability to transfer large files reliably.

To overcome these limitations, Open Exeter is developing its own submission tool using elements of the SWORD protocol combined with the leading research data transfer service Globus. SWORD allows us to query the repository to determine which collection the user is allowed to submit to and what sort of metadata is needed. The tool then gathers the metadata and data locations from the user before scheduling transfer of this to the repository. This method works irrespective of the volume of data and its location whilst remaining secure, fast and resilient since if a transfer fails it can be restarted automatically from the point of failure. Globus provides a unique reference number to track progress and determine completion allowing subsequent submission via batch import to the repository.

Using these technologies, Open Exeter is working toward a solution that will allow researchers to upload their data quickly and securely and will be giving a demonstration of its prototype.