Cloud-like services to improve the preservation of digital cultural heritage

For the benefit of cultural organisations the National Library of Scotland, working with Edinburgh Parallel Computing Centre (EPCC) and with the support of the National Galleries of Scotland and the Digital Preservation Coalition will explore the potential of EUDAT cloud-like services to preserve European digital cultural heritage.

The National Library of Scotland and the National Galleries of Scotland are two of the largest custodians of physical and digital cultural heritage in Scotland covering art, literature, the written and spoken word, music, moving image, the web and digital archives. Each has strong digital expertise and we work together to improve the storage of digital culture to ensure that we preserve access to it.

The Scientific Challenge

Cultural organisations need to preserve access to an increasing amount of digital content that they are creating and acquiring. For example the National Library of Scotland expects its data to grow 10 times over 10 years. This growth increases the strain on the core preservation requirements to store data in multiple geographic locations (cost of setting up more data centres), and to check if data changes over time (costs of increased computing power/time).

High level studies suggest that traditional cloud services offer no net benefit for large volumes of data (100s of TB) that require on-going access to undertake preservation actions. There is little openly published information that describes or quantifies the practical limits and costs of using cloudlike services. For example how long will it take to transfer data? Is transfer and data monitoring scalable? What additional tools and services are required to automate the process?

Cloudy Culture will use EUDAT to hold a safe preservation copy of data to allow locally held access copies to be repaired if they change over time. For this reason access to the copy at EUDAT will be restricted to those few people who are undertaking preservation actions on the data. However the local copy of the data, mainly digitised collections, is freely and openly available via www.nls.uk where the audience size is millions of visitor sessions per year.

Who benefits and how?

The pilot will benefit the targeted digital preservation and cultural heritage communities in four main ways:

- Improve the understanding of costs and benefits for this type of storage by providing reusable, open, real-world data about the performance of EUDAT services in a preservation context that can be transposed to other cloud-like services.
- Understand the viability of cloud-like services for large amounts of data by answering questions such as: Can every part of the chain between the data supplier and the service handle the processing of hundreds of terabytes or petabytes of data each year? Is it viable to transfer data over a 100Mbps or 1Gbps network connection or does the data need to be physically posted? How scalable is the computing power for preservation actions?
- Improve tools for data management by improving and documenting the use of existing EUDAT and other tools to increase the automation of data management so they are useful to other organisations using Windows or Linux.
- Position EUDAT as a digital preservation option by helping EUDAT as part of a federated European infrastructure develop business models for the sustainable preservation of digital cultural heritage beyond its current period of funding.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Technical Implementation

The Cloudy Culture pilot has 4 technical aims. Work has started on the first two that is reported here:

1. To transfer data to EUDAT services over the internet: The pilot has implemented a transfer process using iCommands on an Ubuntu VM running on a Windows host at the National Library of Scotland. The Library is the Cloudy Culture partner that supplies data to EUDAT.
2. To fixity check the data: Cloudy Culture has implemented fixity checking within iCommands/iRODS at the point of transfer using the iput commands –K switch to force the creation and verification of file fixity at either ends of the transfer process. Fixity is independently confirmed by comparing the iCAT record for the file against checksum values generated by the Library at an earlier point in the workflow not using iCommands/iRODS.
3. To run arbitrary software in EUDAT services against the data stored in EUDAT.
4. To transfer data out of EUDAT services over the internet.

Preliminary Results

Cloudy Culture has transferred over 118 batches of files consisting of 924 thousand files, 25TB of data to EUDAT/EPCC. 96 of the batches have been for live data, as opposed to test or debugging data. Only the 96 batches are considered below. They are typically 400GB in size, but for various tests some batches have been as small as 0.08GB. In summary:

- Of 96 batches transferred 1 was interrupted by a power cut at the Library, 5 were interrupted by a loss of connection with EUDAT/EPCC.
- Most batches transferred at a speed of 80 to 100 megabits per second, lower than expected for a 1Gbps connection between the Library and EUDAT/EPCC. The cause of the limited transfer rate has yet to be identified.
- Using the single-threaded or multi-threaded option in the iCommands iput command made no distinct difference to transfer speeds.
- There is a strong relationship (R^2= 0.9264) between the mean file size of the batch and the transfer speed. For example a 400GB batch with a mean file size of around 0.01MB takes approximately 5000 times longer to transfer than a 400GB with a mean file size of 300MB. This means transferring small files without putting them in a wrapper format is highly inefficient. The cause of this penalty for small files has yet to be identified but research indicates that it is not driven by the overheads of querying iCat or generating checksums during the transfer process, which account for a performance hit of around 20%.
- The maximum theoretical transfer rate is 410TB per year based on the batch with the fastest transfer rate.

Of the more than 924 thousand files transferred to EUDAT/EPCC and ingested into iRODs iCat catalogue none has a different checksum value after transfer.

Two files were transferred to iRODs, but did not have a checksum value added to their iCat entry. The cause of this has not been determined and will not be investigated further but highlights a rare event that needs to be intercepted if a) iCat is used to confirm the initial transfer b) iCat is used to check file fixity over time.

Initial timing tests for doing periodic, e.g. annual, fixity checking indicate that theoretically it would take 127 hours to check 40TB of files with large mean file sizes (300MB+), but 35 years to check 40TB of files with small mean file sizes (2-3KB). For the actual 25TB of files currently in EUDAT/EPCC fixity checking is estimated to take 123 hours.

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

Contacts

- Lee Hibberd, National Library of Scotland, L.Hibberd(at)NLS.UK
- Pascal Dugenie, CINES, dugenie(at)gmail.com

Further Information

Scientific Challenges behind the Pilot
Learn more about the EUDAT/Cloudy Culture collaboration here

Read more

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.