DRS Transfer Times

Jim Katz

**Abstract**

Calculates the time and network requirements for preparing and uploading batches of works to Harvards DRS using their provided BatchBuilder and transfer facilities.

**References**

Github : ([https://BuddhistDigitalResourceCenter/drs-deposit](https://buddhistdigitalresourcecenter/drs-deposit)) Specifically:

output/

[1]RS3Archives2017-12-06.xlsx Catalog of all works on RS3 server, with aggregate statistics

[2]LogTime.xlsx Observations of file copying time

[3]timeBatchBuild.xlsx Observations of batchBuilder times

docs/

[4] BDRC Deposit Workflow Plan.doc

[5] DRS2 - Tibetan Buddhist Resource Center Report - Tricia\_Patterson-20160317.docx

# Overview

Building and depositing to DRS has three principal time-consuming steps, regardless of BDRC implementation. These are:

1. Copying the images into the batch structure. This is a copy from one BDRC resource to another, internal.
2. Creating the batches.
3. Transferring the resulting batches.

# Basic statistics

## Quantities

The top level quantities and limits in this discussion are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **BDRC Description** | **DRS Unit** | **Count** |
| Work | A collection of volumes | A collection of batches | 12000 |
| Volumes | A collection of pages, a subset of a work. | Up to 30 volumes can be in a batch |  |
| Images | One imaged page. | Object | 13,670,868 consuming around 5.3Tb |

## Capacities

|  |  |  |
| --- | --- | --- |
| **Unit** | **Owner** | **Value** |
| Disk storage | BDRC | 9Tb dedicated drive |
| Upload limit | HUL DRS | 250000 images/day |
| Transfer speed | BDRC and HUL | TBD |

# Image copy

Every image sent to DRS needs to be copied to a buffer from which it is sent. Batch building does not support sending the batch from its original file location. The copy file occurs on a mac which has both the source and the destination connected through AFP. Measurements[2] have observed throughput of 290 – 300 kb/sec. At this rate, the file copy portion will consume 6 days to copy all the images. (note this 6 days is included in the running time of all the batch builders – it is not separate.

# Batch creation

The most significant components of these are:

* 1. Calling a DRS web service to derive and detch information about an image
  2. Assembling all the images into one xml file.

Measurements on a small sample of files, on a dual core MacOs machine[3] showed batch building process files rates of around 10 files/sec. Actual throughput increases for large batches (more files /sec).

At this rate, the batch building steps would consume 15 days to process.

A second test was done processing 13 batches simultaneously. See sheet ‘batches113’ in [1] . This tripled the throughput, giving an estimated batch building elapsed time of 5 days.

# Transfer

The time to transfer the total BDRC archive is limited by three factors:

1. Transfer rate
2. Limits on daily amount transferred
3. HUL processing time per batch

See [1],Worksheet “Transfer Rate”: If we could transfer the maximum allowed number of files per day, around 55 days would be required.

Network speed is required transfer the maximum file count: around 9Mbits/sec.

We will probably not attain this throughput regularly, as we’re transferring over a VPN. We will measure when we transfer to HUL’s QA system, and advise from there.