

# Digital Preservation at Oxford and Cambridge

A collaborative research project to evaluate and provide sustainable recommendations for our digital preservation programmes

## Validating half a million TIFF files. Part One.

Posted on **2 May, 2017** by **James Mooney**

*Oxford Technical Fellow, James, reports on the validation work he is doing with JHOVE and DPF Manager in Part One of this blog series on validation tools for auditing the Polonsky Digitization Project's TIFF files.*

In 2013, The Bodleian Libraries of the University of Oxford and the Biblioteca Apostolica Vaticana (Vatican Library) joined efforts in a landmark digitization project. The aim was to open up their repositories of ancient texts including Hebrew manuscripts, Greek manuscripts, and incunabula, or 15th-century printed books. The goal was to digitize over one and half million pages. All of this was made possible by funding from the Polonsky Foundation.

As part of our own Polonsky funded project, we have been preparing the ground to validate over half a million TIFF files which have been created from digitization work here at Oxford.

Many in the Digital Preservation field have already written articles and blogs on the tools available for validating TIFF files, [Yvonne Tunnat](#) (from ZBW Leibniz Information Centre for Economics) wrote a [blog](#) for the [Open Preservation Foundation](#) regarding the tools. I also had the pleasure of hearing from Yvonne and Michelle Lindlar (from TIB Leibniz Information Centre for Science and Technology)

talk at [IDCC 2017](#) conference on this very subject in more detail when discussing JHOVE in their talk, [How Valid Is Your Validation? A Closer Look Behind The Curtain Of JHOVE](#)



*The go-to validator for TIFF files?*

### **Preparation for validation**

In order to validate the master TIFF files, firstly we needed to retrieve these from our tape storage system; fortunately around two-thirds of the images had already been restored to spinning disk storage as part of another internal project. When the master TIFF files were written to tape this included MD5 hashes of the files, so as part of this validation work we will confirm the fixity of all the files. Our network storage system had plenty of room to accommodate all the required files, so we began auditing what still needed to be recovered.

Whilst the auditing and retrieval was progressing, I set about investigating validating a sample set of master TIFF files using both [JHOVE](#) and [DPF Manager](#) to get an estimate on the time it would take to process the approximate 50 TB of files. I was also interested to compare the results of both tools when faced with invalid or corrupted sample sets of files.

We setup a new virtual machine server in order to carry out the validation workload; this allowed us to scale this machine's performance as required. Both validation tools were going to be run on a RedHat Linux environment and both would be run from the command line.

It quickly became clear that JHOVE was going to be able to validate the TIFF files a lot quicker than DPF Manager. If DPF Manager is being used as part of one of your workflows, you may not have noticed any real-time penalty when processing small numbers of files, however with a large batch, the time difference with the two tools was noticeable.



*Potential alternative for TIFF validation?*

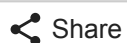
During the testing I noticed there were several issues with DPF Manager, including the lack of being able to specify the number of threads the process could use, which I suspect resulted in the poor initial performance. I dutifully reported the bug to the [DPF community GitHub](#) and was pleased to see an almost instant response stating that it would be resolved in the next monthly release. I do love Open Source projects, and I think this highlights the importance of those using the tools being responsible for improving them. Without community engagement, these projects are liable to run out of steam and slowly die.

I'm going to reserve judgement on the tools until the next release of DPF Manager. We will then also be in a position to report back on our findings from this validation case study. *So check back with our blog for Part Two.*

I would be interested to hear from anyone else who might have been faced with validating large batches of files, what tools are you using? what challenges have you faced? Do let me know!

---

#### SHARE THIS:



This entry was posted in [digital preservation](#), [digitisation](#), [digitization](#), [technology](#), [tools](#) and tagged [community](#), [dpf manager](#), [github](#), [jhove](#), [open source](#), [tiff](#), [tool](#), [validation](#) by [James Mooney](#). Bookmark the [permalink](#) [<http://www.dpoc.ac.uk/2017/05/02/validating-half-a-million-tiff-files-part-one/>] .

Pingback: [Validating half a million TIFF files. Part Two. | Digital Preservation at Oxford and Cambridge](#)

**Tyson**

on **8 June, 2017 at 06:15** said:

I adore it when folks come together and share views, great site, keep it up.

This site uses Akismet to reduce spam. [Learn how your comment data is processed.](#)