Digital Object Curation At Scale

Tom Creighton, FamilySearch (USA); Jonathan Tilbury, Tessella plc (UK); and Mark Evans, Tessella Inc., (USA)

Abstract

FamilySearch acquires digital images of records having genealogical significance from all over the world. These records are typically hand written and are often in the form of volumes of bound forms housed in small archives, parish and church record repositories, and in some cases government records archives. FamilySearch has been involved in capturing images of these records from all over the world since the 1930s. Until recently these images were captured using film cameras and preserved mostly on microfilm. In recent years the film cameras have been replaced with high resolution digital cameras.

Over the years FamilySearch have placed more than 3.3 million rolls of microfilm in their records vault drilled into the side of a mountain in the canyons above Salt Lake City, UT, USA. They are now in the midst of an aggressive digitization effort with the intent to create digital copies of all the images from all microfilm rolls. They expect to complete this digitization effort by end of year 2020.

In addition to the microfilm scanning project, FamilySearch are collecting genealogical records in digital form at a rate of greater than 127 million images per year. These are gathered using high resolution digital cameras in a purpose-built imaging workstation. The workstations can easily be disassembled, moved to a new area and reassembled to support digitization efforts at various areas around the world. FamilySearch currently ingest approximately 33 MB of digital image data from these distributed capture stations. By the year 2020 they expect this capture rate to more than double due to increased resolution and quality of the images, as well as due to the conversion to color imaging. It is likely to even be higher than that since more capture stations are very likely to be commissioned in that same time frame.

The volume of data and its rate of ingest place large constraints on both the processes involved in digital curation and the design of infrastructure to support those processes. By the end of 2011, FamilySearch expects to be managing more than 403 million images. The preservation copies of these images will consume nearly 8 PB of storage for just one copy each. The project must support high scalability in several different dimensions. For example, the total storage demands require consideration of inexpensive media, while requiring maintenance of low error loss. Multiple copies of the digital objects will increase the storage requirements even more. In addition, the system must handle a very high data throughput rate in order to sustain ingest rates of greater than 100 million images per year.

In this document we describe many of the details regarding how the scale issues noted above influenced FamilySearch's digital curation processes, as well as some of the infrastructure design considerations that were made to support them. We also describe how end user access is provided and why FamilySearch chose to separate end user access from preservation repository access.

