Overview

This document describes the current state of Large File support in FEDORA repository architectures, and analyzes situations which require more homogenous support and integration of these large files.

Introduction

Since 2004, the RUcore development team at Rutgers University Libraries has developed and implemented an institutional repository centered around the Flexible Extensible Digital Object Repository Architecture (presently known as FEDORA Commons). Using this architecture, we have developed a repository which can adapt to a wide variety of file formats and object types, and provide a location to digital preserve these objects for the long term. Through this system, and through the use of FEDORA's object management capabilities, our repository currently manages 15,341 different objects, that can be categorized into one of the following major object archetypes:

- **Still Photographs**, either scanned (digital surrogate) or originating from a digital source (born digital)
- **Books and documents**, of varying page lengths. These too can either be digital surrogates or born digital. Optical character recognition is supported in most practical cases
- **Audio and Video recordings**
- Support for **datasets** is currently in the early stages of development.

Our experience with FEDORA and implementing it for institutional use has also brought it with keen awareness of the architecture's limitations. Chief among the challenges we face with FEDORA is the issue of handling large files that exceed 2 Gigabytes (GB) in size. Until very recently, with the release of FEDORA 3.2, it was simply not possible for FEDORA to directly manage objects exceeding this size, and even now, the current method for handling such objects is through the use of a managed external datastream.

While the use of managed-external datastreams is an effective short-term solution, it precludes our ability to process these larger files in ways similar to our internally-managed objects. A significant amount of manual processing must go into the ingest of every externally-managed object, including the generation of digital signatures, manual creation of XML to document file characteristics and some technical metadata, and transcoding these files into web-ready formats; all processes which are handled automatically for still images, documents, and smaller audio files by an internal pipeline process.

We understand that FEDORA developers feel that this is an ideal solution for handling such objects, however there are instances where a more practical, homogenous solution for handling large files in the same way as smaller objects be incorporated into FEDORA, whether it be by streamlining the ingest process or using methods which do not invoke performance penalties for larger objects. This document will outline some of the real-world scenarios we've encountered and how external datastreams impact their use, and will also enumerate some of the future issues we perceive as requiring the need for large files in the not-too-distant future.

Sample instance: Video Objects

Video has by posed the greatest challenge to digital preservationists, and has given us an opportunity to test the true adaptability of RUcore and the underlying Fedora platform. Our greatest challenge was that moving images have an inherent tendency to produce exceptionally large preservation data streams, larger than any other object type that RUcore had endeavored to ingest in the past. Typically, an uncompressed, full-frame video file will take approximately 10 GB of disk space for each half hour of recorded Standard Definition video, with HD content far exceeding this estimate. The large size of these videos and FEDORA's inability to handle them as a directly-managed datastream caused significant hurdles in our development process.

An early interim solution involving segmented archival data streams that could later be reconstituted to retrieve the archival master video file was tried, but proven to be impractical and risky. Fortunately, current versions of Fedora support external data streams, and we have implemented this architecture into the default content model for video.

The CM ultimately implemented consists of an externally managed Archival Datastream, which can take the form of an uncompressed AVI file for video converted from analog formats, or a capture of the digital master content for born-digital video (such as video captured in DVCAM, MPEG-2, AVC MPEG-4, etc.), bundled into a tar file. Depending on the nature and condition of the archival master, high-quality derivative "helper files" such as a high-bitrate H.264 video stream may also bundled into the tar archive to potentially assist future transcoding efforts, and act as a "sanity check" to compare against the original master.

Sample instance: Audio Objects

In late summer 2009, a test ingest of audio objects from The Rutgers Jazz Oral History Project (JOHP) on our development installation of the FEDORA repository revealed that we would encountered large file issues in audio objects as well, in limited instances. JOHP is unique in that it is the first major audio collection we have attempted to ingest into RUcore, and consists of a very large number of segmented WAV files. These WAVs are derivatives of reel-to-reel tapes that were digitized onto CD Audio discs, and their segmentation is based on the limitations of both the tape length and disc capacities.

Prior to ingest, these WAV files are concatenated for continuity. The result is a collection of audio files containing interviews of Jazz musicians, some lasting as long as 8 hours, with WAV files taking up multiple gigabytes of space. We had not anticipated that we would encounter large file support issues for audio. Consequently, we've been forced to delay ingest of this collection while we explore adding a workflow to ingest these audio files as an externally-managed datastream.

Object Types with possible near-to-moderate-term risk for increasing file size

Our digital curation program at Rutgers has given us a great deal of experience with handling file types from both emerging file formats as well as mature structures that are evolving as technology advances. Audio and video are merely giving us a taste of the volume of data that we can expect to see on a regular basis and consequently, will be called upon to preserve digitally.

- **Still Images** - Although initial digital images were of rather low resolutions (8 Megapixels or less), we are now starting to see increasingly higher pixels counts from digital still cameras that have more advanced imaging sensors. At this stage, it is now possible to obtain born digital images that are in the gigapixel range using image compositing and High Dynamic Range (HDR) technologies. Such images will find applications in scientific research where image analysis and feature mapping are crucial.

- **Data Sets** - The preservation of scientific data is a field that RUcore is now entering, and we anticipate that there will be wide variations in how researchers collect and store their data. Some studies may produce relatively small data files, while others could involve massive amounts of numeric data. We would like to see a workflow where we could treat all datasets we ingest equally, not having differentiate between small files which can be ingested as managed FEDORA data streams, and large sets which must be managed externally.