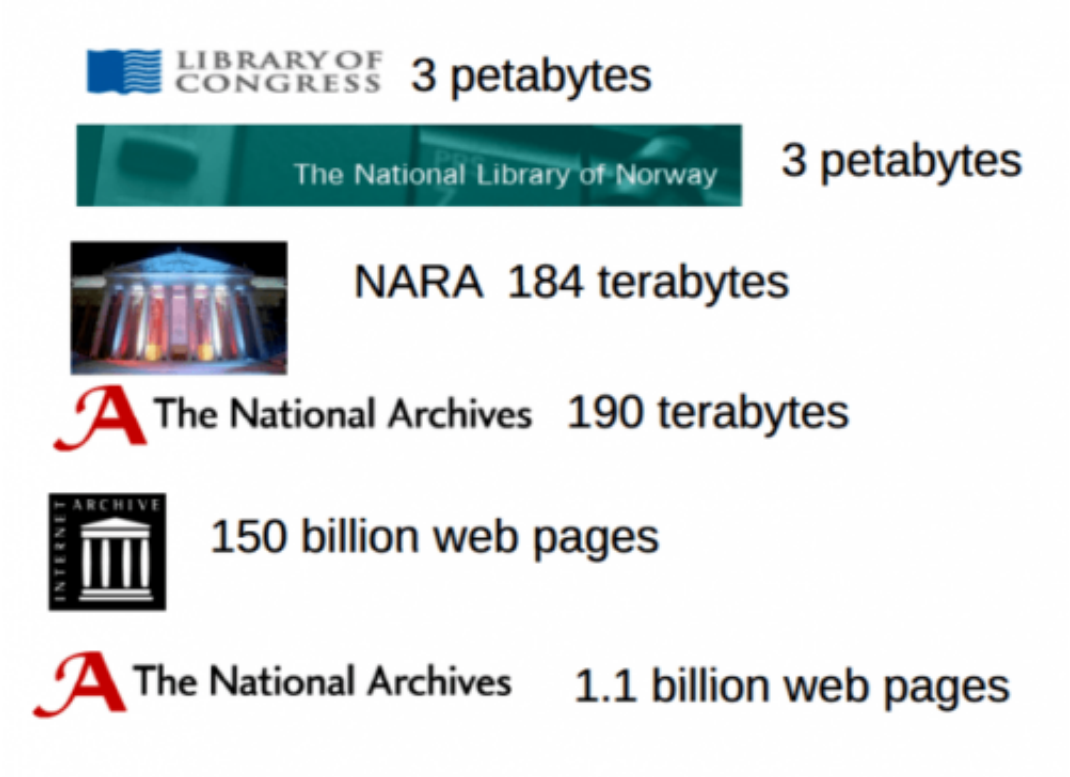
How one of the world's largest archives is managing the move from parchment to pixels

January 16 2015, by David Clipsham



Measuring up: the world’s largest archives. Credit: National Archives

From the Domesday Book to modern government papers, the National Archives' collection of more than 11m historical government and public records is one of the world's largest. It includes paper and parchment, photographs, maps and paintings, but also a vast number of digital records such as archived government websites, emails and social media

posts. Paper may last for thousands of years, but what about the ever-expanding quantity of digital documents?

The National Archives' broad remit under the Public Records Act is to permanently preserve the records of the UK government that have been selected for their historic value.

Our physical records that date back over 1,000 years take up more than 200km of shelving and require delicate conservation work and careful storage. The digital age on the other hand requires little physical space but presents different challenges – how do archivists cope as we move from parchment to pixels?

Build the tools, and they will come

We began to focus on the challenges of digital preservation during the mid-1990s, realising at the time that there was no authoritative and centralised source of information regarding file formats. So we developed PRONOM, a registry of file formats and the applications required to open and read them, and DROID, a freely available open source tool to manage that data and information.

To date, PRONOM contains details on more than 1,000 different file formats including their technical specifications and a reliable method for identifying them using a byte-level analysis of a digital file.

Together, DROID and PRONOM are used by heritage institutions the world over, and have found their way into dedicated archival storage software, and even digital forensics tools used by police and investigators.

Kilometres of these now fit in a few cubic inches of digital storage. Credit: merlin1487, CC BY-NC-ND

Parsimonious preservation

The approach to digital preservation that we've developed we call parsimonious preservation, which can essentially be distilled down to two principles:

- Understand what you have got
- Keep it safe

To keep our digital records safe, we have built our own in-house Digital Records Infrastructure, with various features. For example, several layers of anti-virus scanning to ensure we're not exposing our records to corruption from malicious programs. File fixity checking, to ensure that any digital object received has not been altered or lost bits and become

corrupted since being archived. Properly identifying a file, via DROID and PRONOM, so that the filetype and how it can be read is properly recorded. And metadata validation, to ensure that the information held that describes a record in the archive matches the record. When anything is to be released under the Public Records Act, we do so through our online public catalogue, Discovery.

Since 2003 the National Archives has maintained and expanded the UK government web archive. To date we have captured around 100 terabytes of material, and this is growing by roughly 1.5-2 terabytes per quarter. The oldest website captured (the Ministry of Defence) dates back to 1996, and in a typical month we will capture around 20m unique URLs. The web archive is important to provide continued access to historical government information that has been released online as various departmental or government agency websites are renamed, moved, merged, or closed.

Now that social media has become more prevalent as a communication tool of government, we've expanded our social media archiving project in order to capture and preserve Twitter and YouTube feeds of government departments. Presently we capture the output of 67 UK government Twitter accounts, and up until September 2013 had captured over 65,000 tweets and around 7,000 YouTube videos.

Our digital archive currently has a potential capacity of 13.7 petabytes – almost 14m gigabytes – of which the current archives have used one petabyte. The archive system is built in a modular fashion, which means software or hardware components can be added or replaced as technology improves, adding capacity or improving processing power – an archive fit for the future. This is just as well, as the next few years will see a significant increase in the volume of digital records generated by the UK government that will require archiving for the future.