Norway's petabyte plan: Store everything ever published in a 1,000-year archive

From ancient manuscripts to movies, the National Library of Norway wants to put it all online for the public.

Written by **Stig Øyvann,** Contributor

Oct. 4, 2018 at 2:38 a.m. PT

In the far north of Norway, near the Arctic Circle, experts at the National Library of Norway's (NLN) secure storage facility are in the process of implementing an astonishing plan.

/ ar + vr

I replaced my boring workouts with Meta Quest's Supernatural app, and can't imagine going back

This Finnish startup's new VR headset rivals Apple's Vision Pro - and business users will love it

Meta's $500 Quest 3 is the mainstream VR headset I've been waiting for, and it delivers

I tried Apple Vision Pro and it's far ahead of where I expected

The best VR headsets right now (and they're not just from Meta)

They aim to digitize everything ever published in Norway: books, newspapers, manuscripts, posters, photos, movies, broadcasts, and maps, as well as all websites on the Norwegian .no domain.

Their work has been going on for the past 12 years and will take 30 years to complete by current estimations.

At the moment, the library has more than 540,000 books and over 2,000,000 newspapers in its archive. These have been mass-scanned and OCR-processed before being stored, so all the content in the library is free-text searchable.

SEE: Sensor'd enterprise: IoT, ML, and big data (ZDNet special report) | Download the report as a PDF (TechRepublic)

As of early September, the collection amounted to 8.1 petabytes of data and is growing by between five terabytes and 10 terabytes every day, Svein Arne Solbakk, department director for digital library development at the NLN, tells ZDNet.

NLN's mandate isn't just long-term safe storage. It is also making its archives available for the public, so it needs online storage for publishing the collection.

"Just to be able to handle the large amounts of data, we must have it online. If I get a PDF file from a newspaper, I know this format won't last for a thousand years. I'll have to convert it to a modern format, probably several times during those thousand years," Solbakk says.

He illustrates this point by explaining that they've already had to complete their first large-scale format conversion, involving 50 million image files. This process took 10 servers three months of 24/7 processing to complete, even though the files were stored on hard disks.

Furthermore, given the relatively short life of hard disks, the NLN's approach is to have a rolling program of disk replacement, swapping out entire disk cabinets when they reach their expected lifespan of five years.

SEE: Digital transformation: A CXO's guide (ZDNet special report) | Download the report as a PDF (TechRepublic)

In addition, the NLN stores everything in triplicate. One copy is on hard disk, with two more copies on tape. The tape storage is an archive system based on Oracle SAM-FS, so it's not a traditional tape backup system.

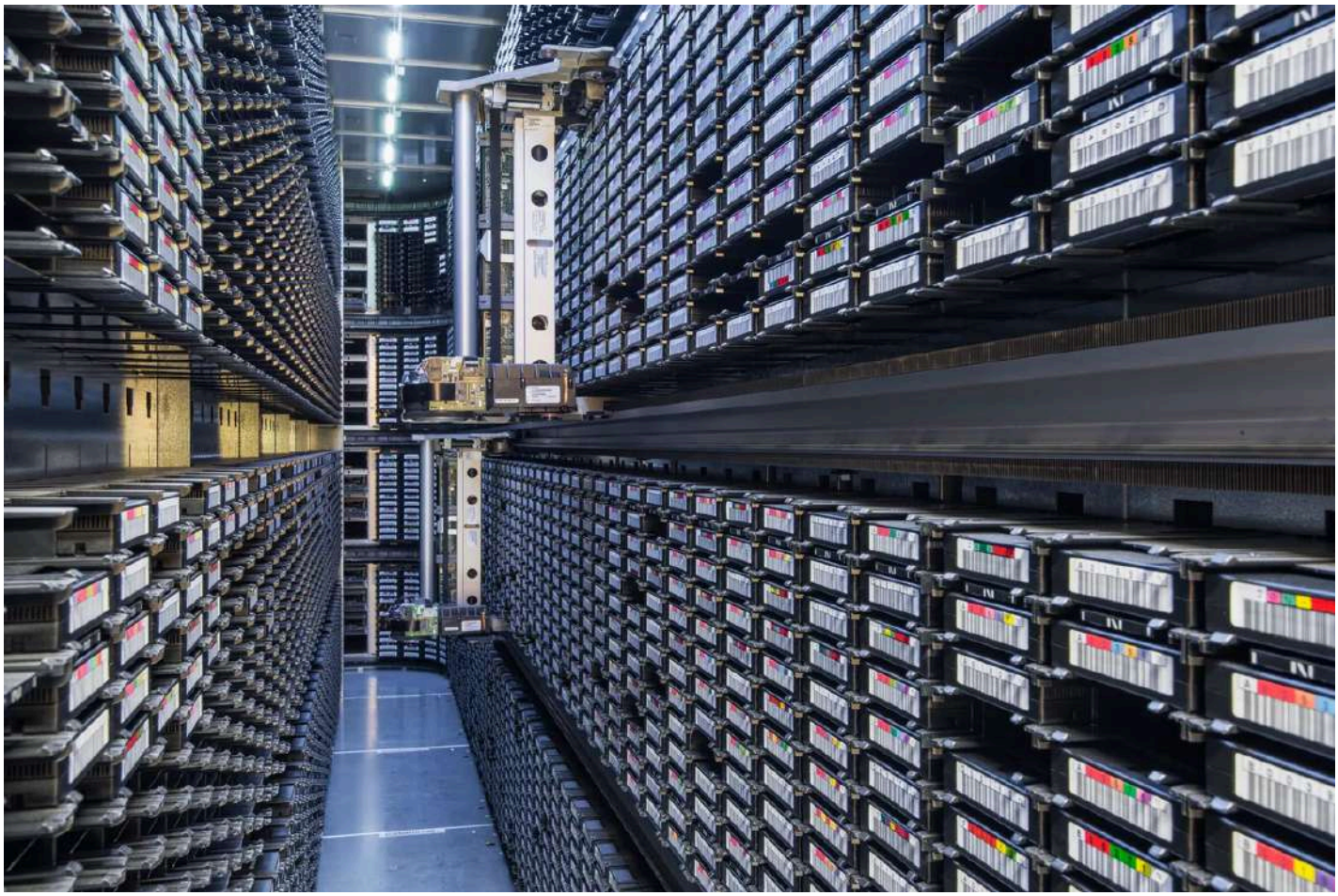
"When we're talking petabytes, we can't talk about backup. A petabyte restore from tape would take weeks," Solbakk says. Thus, the NLN's system is more of a storage-virtualization approach that is currently handling more than 24 petabytes in total.

Some 83 percent of all books and 40 percent of all newspaper pages have been digitized. In addition, the NLN is among several other projects currently working on scanning 100,000 radio broadcast tapes before the tape players needed for the job disappear for good. It's easy to be impressed by the NLN's ambition.

"We are ambitious, but it's very important to document the present for the future," Solbakk concluded.

National Library of Norway's digital collection, September 2018

- 2,000,000 newspapers, about 40,000,000 pages
- 540,000 books, about 80,000,000 pages
- 700,000 pages of manuscripts and music manuscripts
- 1,300,000 photos
- 1,400,000 hours of broadcast radio
- 950,000 hours of broadcast TV
- 55,000 units of music
- 16,000 units of movies/video
- 24,800,000,000 web pages



The tape storage is an archive system based on Oracle SAM-FS, so it's not a traditional tape backup system.

Image: Nasjonalbiblioteket/Jan Inge Larsen