# Video on Data Sharing
# Originated by Robbie Laird (WHOI) on Jan 7, 2014

**From Robbie Laird (WHOI) on Jan 7, 2014**

Hi

Probably we are not supposed to send jokes on this mailing list, but this is truly educational, as well as funny. (sound is not essential, but it's better with it.)

> http://www.youtube.com/watch?v=N2zK3sAtr-4#!
> <http://www.youtube.com/watch?v=N2zK3sAtr-4#%21>
>
> Apologies to those who have already seen it.

Robbie Laird
WHOI/SSSG

---

**Reply From: Steven Roberts (UAF) on Tue, 7 Jan 2014**

According to the following article these issues with data sharing are sadly the norm:

http://blogs.smithsonianmag.com/science/2013/12/the-vast-majority-of-raw-data-from-old-scientific-studies-may-now-be-missing/

Steve Roberts
UAF

---

**Reply From: Webb Pinner on Tue, 7 Jan 2014**

Along the same thread but hitting a little closer to home…
I came across a paper written by David Fischman from the NGDC talking about the costs of data stewardship compared to the costs of losing the data and having to recollect. He focuses mainly on the multibeam data submitted to NGDC by NOAA and UNOLS. It was a pretty shocking number. Last year David gave me permission to re-post his paper on my website. Here's the link if anyone wants to read it:
http://www.oceandatarat.org/?p=783

Cheers,
- Webb

---

**Reply From: "David O'Gorman" (OSU) on Tue, 7 Jan 2014**

Steve,
That's a pretty sad result. That reminds me of a really great article I read a while back about the pioneer probes slowing down gradually as they transited the solar system ("Pioneer anomaly" http://en.wikipedia.org/wiki/Pioneer_anomaly).

The short story is that the pioneer probes were affected by a slow acceleration which added up to 1kph over a period of ten years. After a wide variety of other explanations were suggested a team of investigators determined that the acceleration was due to heat from the batteries reflecting off of the back of the probes.

The longer story is here:
http://spectrum.ieee.org/aerospace/astrophysics/finding-the-source-of-the-pioneer-anomaly
And involves a team of scientists going so far as to re-analyze as much of the original data as they could lay their hands on to precisely determine the magnitude and nature of the unexplained acceleration.

Highlights include:

"As luck would have it, most of the Pioneer 10 and 11 telemetry data had been saved and were available for study...a contractor and former Pioneer team member at NASA Ames Research Center, had been informally preserving all the Pioneer data he could get his hands on. Kellogg already had nearly all of the two probes™ master data records, binary data files that contained all the Pioneers™ science and housekeeping data.

Kellogg had taken care to copy those records, which in total took up just 40 gigabytes of space, from soon-to-be obsolete magneto-optical discs to a laptop hard drive. ... working on new software that could extract useful information from the master data records without the need for an old, decommissioned mainframe.

... we were able to find additional files on the hard drives of JPL navigators™ computers and the archives of the National Space Science Data Center. We even found magnetic tapes stuffed in cardboard boxes under a staircase at JPL. Some of the files were in a rather sorry state, corrupted while they were converted from one storage format to another over the span of three decades."

Reading this article really drove home the importance of good data archiving methods for me.
Dave

David O'Gorman
Marine Technician Superintendent OSU CEOAS
STARC

---

**Reply From: "Cohoe, Dave" (USCG) on Thu, 9 Jan 2014**

Would it be reasonable to require long-term data storage plans in research proposals?

Regards,
Dave

Dave Cohoe, CISSP, ITIL
Polar Science Systems
Base Seattle C4IT
U.S. Coast Guard

---

**Reply From: Ethan Gold (Ocean Exploration Trust) on Thu, 09 Jan 2014**

I certainly think so, especially if it is expected that the data will be publicly available. Just yesterday I was talking to our video producer about budgeting for periodic conversion of video footage onto whatever the plausible archival media of the decade might be. Otherwise all that fantastic historical data is no more than ballast.

-E

---

**Reply From: Alexander Shor (U. Hawaii) on Thu, 9 Jan 2014**

Ethan, Dave and others:

The NSF OCE data policy can be found here: http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf

It covers most data (samples and digital) collected during NSF-funded projects.
It provides for archiving of data in public repositories where they do exist.
Some specific NSF programs have additional requirements.
NOAA also has a published data policy, as does ONR.
There have been discussions in the past at RVTEC meetings, but I have not attended recently, so not sure how recently.
Jim Holik can probably update you.

Sandy

Alexander (Sandy) Shor
Associate Dean for Research, SOEST
University of Hawaii at Manoa
1680 East-West Road, POST 802
Honolulu, HI 96822 USA

**Reply From: Dale Chayes (LDEO) on Thu, 9 Jan 2014**

NSF started a requiring a (more or less) robust data plan at least a year ago.

-Dale

---

**Reply From: “Holik, James S" (NSF) on Thu, 9 Jan 2014**

A few thoughts:

NSF will not even accept a proposal without a Data Management Plan. What that means varies by program but all proposals to NSF need to address this issue. As is stated in the OCE Data policy, all underway data collected on ships in the Academic are ingested and cataloged through the R2R gateway and ultimately sent to the appropriate long-term archives run by NOAA. For all data collected at sea and not part of the R2R program, the PI has the responsibility to do the same thing. As the types of data change (video, etc) I feel confident that the data fanatics at R2R et al, are developing ways to collect, document and archive it.

Jim

Jim Holik
National Science Foundation
Program Director
Ocean Instrumentation and Technical Services

---

**Reply From: Webb Pinner on Thu, 9 Jan 2014**

NOAA has also been working on a video management/archival policy with NOAA Central Library (NCL). Their main challenge has been keeping up with the storage requirements and developing standards for acceptable video formats. When last I spoke with them the quicktime (.mov) format was the preferred wrapper recommended for archival.

As Jim said the types of data are changing, even just within the video world. In the past things were based in physical media like Beta, DVCAM, DVD, HDCAM and Blu-ray. These were the formats used by Alvin, Jason, Hercules/Argus and others. It is now getting harder to support physical media as the consumer and broadcast video worlds have moved to file based recording systems. In the file-based world I'm seeing Apple's ProRes422 as the dominate codec. ProRes422 comes in a .mov wrapper so it is a support by NCL as an archive-appropriate format. It is currently used by the Okeanos Explorer, Neptune Canada, SOI, the new Alvin and is planned for the R/V Sikuliaq.

The ProRes422 codec supports several bitrates for HD video ranging from 50-260Mbit/s. These bitrates translates into 25-125GB/hr for space requirements. If you extrapolate these data rates out (2 feeds, 100Mbit/s codec, 24 hours) you're looking at ~2.5TB/day. So for a 3 week cruise with 12 days worth of ROV dives the storage requirement will be ~30TB.

I bring this up only to help frame the challenge. Storage is absolutely getting cheaper by the day but extrapolating a 30TB storage requirement for just one cruise on one vessel with an ROV get's big and expensive fast. The video is so expensive to collect that it is my option that anything less than hardware-based RAID storage is just too risky to attempt (vs the cheap USB hard drives). The Okeanos uses my preferred setup of dual RAID systems which doubled the infrastructure costs but protected the video data from any equipment failure short of the rack-room catching fire (and I've tried to get one of the arrays moved to a different part of the ship).

If the community does what to archive full-resolution video it is not unrealistic that petabyte storage solution will be required in the next 5 years. A more cost effective solution might be cloud storage options like what's offered by Amazon. I believe this is what the NOAA satellite folks are using.

I would like to bring up the issue of retrieval. Ever tried to copy 1 terabyte of data to a USB2 hard drive? USB2 is 480Mbit but with overhead it's more like 25MB/s… so 1TB can take up to 11 HOOURS! So 30TB could take ~13 DAYS!! USB3 is advertises as 10x faster but now disk read/write speeds come into the equation. Long story short: start those science copies early, scientist may want to include sizable budgets for purchasing data storage and there will most likely aways be a place in this world for sneakerNet.

So how will future users find and retrieve terabytes of video data via the Internet… the challenge awaits.

Cheers,
- Webbb

---

**Reply From: "McGillivary, Philip A CIV" on Thu, 9 Jan 2014**

I have addressed some issues of archiving video from UAS, and raised the potential for storage (by institutions) using Google's cloud storage.
The data are in my SCOAR presentation from this past June, http://www.unols.org/meetings/2013/201306sco/201306scoap07.pdf slides 51-54. I would note the prices quoted are for uncompressed video; so whatever compression rate you are using or wish to use should be factored in.

Webb raises many important points, and a decision by organizations whether they want to deal with the infrastructure or 'farm it out' as with Google or Amazon is bound to be a topic of discussion soon. We will be flying a number of unmanned aircraft on HEALY

again this summer. And places like Norway are moving ahead to install within the next three years a complete coastal surveillance network of unmanned aircraft which will collectively be producing an awful lot of video data. And they are thinking now how to deal with that, and very likely we will find ourselves in a similar position at some point.

Dr. Phil McGillivary

---

**Reply from Ethan Gold (Ocean Exploration Trust) on Thu, 09 Jan 2014**

Quoted for truth.
Indeed, we've had to limit what we can promise during a cruise due to the practical constraints of duplicating so much data and the resources it requires. Managing expectations well in advance is key.

Aboard Nautilus we do non-video data copies incrementally and in parallel, up to 4 at a time, during the aggregation step. By the time the ship docks, the promised copies are ready to go, assuming auxiliary human-dependent bits are all in place. This volume is currently modest enough that it can be kept accessible even on desktop storage arrays.

For video, we make our two hard copies (LTO-6) in parallel (one for ship, one for shore), and all other copies or transcodes are produced ashore post-cruise after the sneakernet transfer has occurred. Currently we don't try to keep all the raw on spinning disk.

But all that simply gets us to the point that spawned this thread.
We are definitely talking about cloud storage at this point. The big data companies have solved this well enough that they can likely be trusted with it for... awhile. In the long run it may just be another media 'format', which has to be converted to something else in another 10 years. A decade from now our current ocean of data will seem tiny, but we'll be trying to stuff 4k video into the data cabinet soon enough.

-Ethan

--
---
Ethan Gold
Director of Software & Data Engineering
Ocean Exploration Trust

---

**Reply From: Brent Evers (IRIS) on Fri, 10 Jan 2014**

The unaddressed issue here is that 99% of video 'data' is completely worthless. Its as much 'data' as a meteorologist pointing a camera straight up at the sky and claiming it all as 'data'.

While the availability and ability to store mountains of video is not an unwelcome technological advancement for science, the question of how to actually use (lots of) it has not been solved, at least not within the science community. The same problem exists within the military and how to deal with UAV surveillance data, but there's orders of magnitude more financial bandwidth to deal with the problem there, and a societal motivator (security) to solve the problem also.

Within the science community, who is going to go and 'mine' that video for useful content after the fact? Who is going to pay for that?

Video is an excellent operational tool (ROV ops, safety monitoring, ship to shore inclusion of land based scientists, etc), but for archival purposes, I'm not sure discreet images don't have more value, as they are taken with an intent and purpose of capturing something of actual scientific interest at the moment (or shortly after the moment if one pulls a frame from recently recorded video). I'm not saying that video shouldn't be archived as we may find (or the capability may trickle down) technological solutions to efficiently extracting useful information from mountains of video over time, but I think its wise to not put the cart before the horse and advocate that huge amounts of storage are implicitly needed before a use case is clearly defined.

OE/Webb and other groups have gone a long way in pushing the metadata component of video so that there is some 'map' of what was recorded, but I'm still not convinced that even that metadata will render all that video scientifically useful.

I think the scientific community should also be cautious in 'counting' video data as a measure of success of an experiment, cruise, etc. I see fantastic claims of "we collected X amount of data from this ship or network and its so much more than ever before", and all I can think is - 'yeah, but no-one is ever going to use it' (because the don't know how to extract information from it at a viable value/effort ratio).

Its a tough problem to solve - both the storage and the use issue - but doing so in a cost effective/financially feasible way is going to also require some tough questions, technological development, etc..

Brent

---

Reply From: Val Schmidt (UNH) on Fri, 10 Jan 2014

This is a fantastic and timely discussion! I have 2 cents.

The problem described for video data also exists in the recording of full water column multi beam sonar data. Although there are not many Reson systems in the deep water fleet, full water column data from a Reson system can result in a GB per minute or so. Water column data in Kongsberg systems is down-sampled so the volume is not quite so large given the same water depth, but it is still large enough that most folks are struggling

to sort out if and how to collect it. Interestingly depending on the number of beams and what is retained in the water column data it can require less storage space to record the raw pre-beamformed element level data than the beam formed data and beam form in post-processing. Few systems provide this capability however.

And like video data, as Brent points out, much of the water column data may not be useful, but like video it is impossible to know a priori.

That said, routine collection of water column by the Okeanos Explorer in the past few field seasons has produced some dramatic finds (100's of previously unknown seafloor bubble seeps) and I believe there is all kinds of untapped information about our oceans in those data files.

-Val