OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)

December 2016

Team Members:
Sharon Mesick, Federal Program Manager
Susan Gottfried, Team Lead
Gina Brewer
Donald Collins
Anna Fiolek
Vidhyadhari Gondle
Denise Gordon
Fred Katz
Yuanjie Li
Andrew Navard
John Relph
Brendan Reser
Jeff Rey
David Sallis

Table of Contents

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 2

Executive Summary

The Video Data Management Modernization Initiative was conceived as a series of projects to investigate and test modern methods of video data management. The premise of the investigation was that annotated digital video readily lends itself to online discovery and direct data access methods. The intention was to preserve the video data record for the long term while at the same time provide users with easy, direct access to these important scientific data.

This investigation was conducted over a period of 4 years, and included
- Participation in the NODC-led cloud pilot project
- Participation in the Federal Agency Digital Guidelines Initiative (FADGI) study
- Documentation of best practices including extensible ISO 19115-2 metadata templates
- Completion of automated data load and extraction systems
- Development of a video data discovery and access portal

This work has culminated in successful implementation of a modern, end-to-end video management system that provides direct online access to the complete collection of digital video collected by various deep submergence systems operated from the NOAA Ship *Okeanos Explorer* (from 2010 to present).

While this current iteration benefits a specific, well-managed data set, elements of the project are viewed as extensible to other video data collections. The joint project team continues to work though the broader OER video collection, which pre-dates the *Okeanos Explorer* (2001-2010). These data are thought to be analogous to the large legacy video collections held by many NOAA programs, and will further challenge best practices developed to date.

This document describes the joint OER-National Centers for Environmental Information (NCEI) Video Data Management Modernization Initiative and seeks to both inform and invite conversation with other programs seeking management solutions for legacy video data inventories and/or planned future collections.

I. Background

Following the recommendations in the [Report of the President's Panel for Ocean Exploration](#), NOAA's nascent Ocean Exploration program began a collaborative partnership with the NOAA national data centers, the National Centers for Environmental Information (NCEI), to inform and implement best data management practices for NOAA's ocean exploration data collections.

In the intervening years, NOAA's Office of Ocean Exploration and Research (OER) has sponsored hundreds of ocean exploration missions, and the joint OER-NCEI team has documented, preserved and provided free public access to untold volumes of data. OER-sponsored missions often involve recording environmental data from cameras mounted on manned and unmanned deep submergence systems. In the early years of the program most of these video data were recorded on physical media, copies of which were catalogued and preserved at the NOAA Central Library. The OER Data Management Team developed and implemented the [original Video Data Management System](#) to support user access to these data.

In 2010, OER initiated a new paradigm for ocean exploration aboard the NOAA Ship *Okeanos Explorer*, America's first ship dedicated to exploring the world's ocean. Integrated communications between shipboard and deep submergence systems, coupled with Internet .connectivity, enable real-time data streaming and sharing the excitement of exploration and discovery with online audiences worldwide. Scientists participate in real-time video stream annotation, and videographers aboard ship edit and further annotate video in near real-time. These activities yield a large digital collection of annotated video "clips" at varying resolutions, as well as a standard suite of compiled video products. As a result, OER video data and information products are highly sought after by the scientific, academic, and broadcast journalism communities.

OER and NCEI embarked on a joint effort to leverage the advances in digital video collection and annotation to improve the experience for end-users seeking to access video data after the live dive event. One of the many roles of the OER Data Management Team is to provide assistance to end users to discover and access OER related data if they are unable to find it themselves under OER's data discovery and access tool, [OER Digital Atlas](#), In the case of the full-resolution video from the *Okeanos Explorer* missions, the volume of these data prohibits archive and access to the data

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 4

through traditional methods. As a result, the OER Data Management Team began a Video Data Management Modernization Initiative (VDMMI) project to come up with a solution to not only preserve these valuable video data assets, but to make them discoverable and accessible in a self-service model. This document describes the project and seeks to inform others with video data inventories that need to be preserved.

II. Video Data Management Challenges

Any organization with a backlog of video data will recognize that several challenges exist in managing those data:
- Video data on physical media has very limited access
- Video data on physical media are at risk due to media deterioration
- Video data are a high-volume dataset
- Video data are in high-demand to a wide audience
- Self-service discovery and access methods and tools are not readily available

In the past, video data management at NCEI was based on a physical media model that incorporated both physical media for the original video and computer files for low-resolution proxies. Called the Video Data Management System (VDMS), it relied on the NOAA Central Library's (NCL) cataloging system for discovery, with online data preview of metadata only. If a user wanted to view the video, they would have to make an appointment with a librarian to physically play the media in the NCL Video Lab which housed tape-playback and DVD decks, or make arrangements for the object to be sent to them via Interlibrary Loan. This method only allows one patron at a time to view the video, further limiting access.

Added to these issues are the challenges posed by legacy video originally captured or ingested onto physical media, such as MiniDV tapes or DVDs. The majority of the video data recorded on research cruises funded through the Federally Funded Opportunity program were submitted on consumer-grade tape-based media, such as MiniDV, and ingested into the archives in its original tape format or burned onto optical discs, such as DVDs. However, physical media is much more subject to environmental and internal deterioration, and if not transferred over to spinning disk files within a certain time, will irreversibly degrade.

When the Okeanos Explorer began its explorations, it brought new video management challenges. *Okeanos Explorer* captures its full-resolution video in a ProRes 422 format at 145 Mbps which creates high-volume files that are 65 GB in size per hour. Data compression is a powerful tool that can be used to decrease file size, but too much compression introduces noticeable artifacts and lowers visual quality. In order to be useful to the scientific and broadcast communities, video must be captured and made accessible at high bit rates to allow for high-definition broadcasting and for effective scientific analysis.

Because of its wide appeal among scientists, broadcast journalists, educators, and the general public, underwater video is in high demand. Video data represents the most sought-after data type managed by OER and therefore, its management is a high priority. In order to answer video data requests, the OER Data Management Team would work with the requestor to identify the correct video segments, and transfer them via ftp or hard-drive, depending upon volume. Each cruise may produce up to 10 TB of video data, delivered to the OER Data Management Team on large RAID-array drives. The shipping of these delicate drives became problematic and the volumes of data became unmanageable. Additionally, the team needed an inventory of portable hard-drives to fulfill data requests. Eventually, the team became painfully aware of the need for a user self-service discovery and access tool and no such tool existed.

III. Pilot Project (2013-2014)

In recognition of this broad array of challenges, in 2013, the OER Data Management Team joined a pilot project already underway at NCEI which was investigating alternative storage and access models for large-volume satellite datasets using the Amazon Cloud services. The OER team used video data from *Okeanos Explorer* as a test data set for the pilot project. For the OER team, the pilot project successful outcomes were:

- Best practices for recording and managing video data were defined through active participation in a Federal Agencies Digitization Guidelines Initiative (FADGI) working group (*see Appendix A*)
- A metadata template (which was successfully vetted through NOAA's Metadata Working Group) was developed to provide geospatial, temporal and vertical extents for each segment and multiple access points for previewing video and ordering full-resolution video from deep storage

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 6

- A middleware prototype for a user interface was designed to collect search criteria from the user and return a list of video segments that meet the criteria

A.   Best Practices from Pilot Project

OER adopted best practices for video recording and stewardship for the *Okeanos Explorer* video following the completion of the pilot project:

- *Capture video at the highest quality levels with the lowest level of compression*
  Okeanos Explorer source video from the ROV or from the accompanying camera platform is captured at 1920 x 1080i employing the ProRes 422 video codec inside a QuickTime container format. At 145 Mbps, this format utilizes mild lossy compression but is still considered broadcast quality and also meets the needs of the scientific community for zooming in for species identification.
- *Collect complete metadata starting at the video shoot*
  Standard Operating Procedures aboard the *Okeanos Explorer* capture metadata in the file naming convention, in the header data of the video file, in the submersible sensors recording environmental conditions, and in the communications going on during filming.
- *Generate a high integrity and continuous master timecode*
  On the *Okeanos Explorer*, the ship's clock and all systems are synched to the industry standard Society of Motion Picture and Television Engineers (SMPTE) timecode referenced to Universal Time Code (UTC) and the timecode is used in the file naming convention. The start time and duration of the source video become critical in gleaning metadata from other sources of information.
- *Move video files to stable storage media as soon as possible*
  The *Okeanos Explorer* has video data storage redundancy in several places: on the ship, at a shoreside repository, and on external hard-drives.
- *Select video encoding and wrapper formats that are standardized, well-documented, and commonly supported by downstream applications, now and in the future*
  The ProRes 422 video codec and the QuickTime container format used by the *Okeanos Explorer* were considered the best choices to meet the requirement for standardized, well-documented, and commonly supported codec.
- *Select video formats that can handle complex audio configurations*

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 7

Okeanos Explorer's video has four audio channels available. The voices of the participating scientists, both onboard and on a conference call line, and the voices of the submersible pilots and navigators are captured on two of the four audio channels.

B.   Metadata

The OER pilot team developed an ISO 19115-2 metadata template that provided for all of the elements for search and discovery. Each element that could be indexed was filled with the most descriptive information available, pulled from available sources. A strict video file naming convention identified cruise, date, camera, video quality, and initial annotation of the subject matter. While this made for a long title, the advantage was that the title, which always remained with the video object, gave a quick and accurate depiction of the contents of the video. Keyword extraction routines provided descriptive metadata that helped uniquely identify the content of the video. The continuous SMPTE Timecode and the segment duration embedded in the video file were used to identify a temporal extents of the segment. The submersible sensor data streams were used to determine the geospatial and vertical extents of each video segment based on the timestamp of the beginning of the segment and the segment duration. The metadata record was also built with two distribution sections which allowed for access to the low-resolution, web-streaming quality video for preview purposes and the full-resolution version for broadcasting and scientific analysis.

C.   Prototype Middleware

Linking together the end user with the desired video data was a customized middleware product that would aid discovery and access. Software developers on the team created a geoportal, a customized front end to handle metadata search through the Amazon Web Service infrastructure. This middleware application was a geospatially-aware discovery and access website that would enable users search through a ISO 19115-2 compliant existing metadata catalog, request a copy from Amazon Simple Storage Service (S3) if available, and if the users wanted the full-resolution, an order would be placed to Amazon's cloud storage service, Glacier, for the video data, which would be retrieved and put back on S3 some within 3-4 hours. The middleware would generate a temporary segment request in a local relational database, including the name of the file and the user's email should the file need to be loaded from Glacier to S3. The requestor was alerted when the file was ready for viewing. See Figure 1 for a schematic.
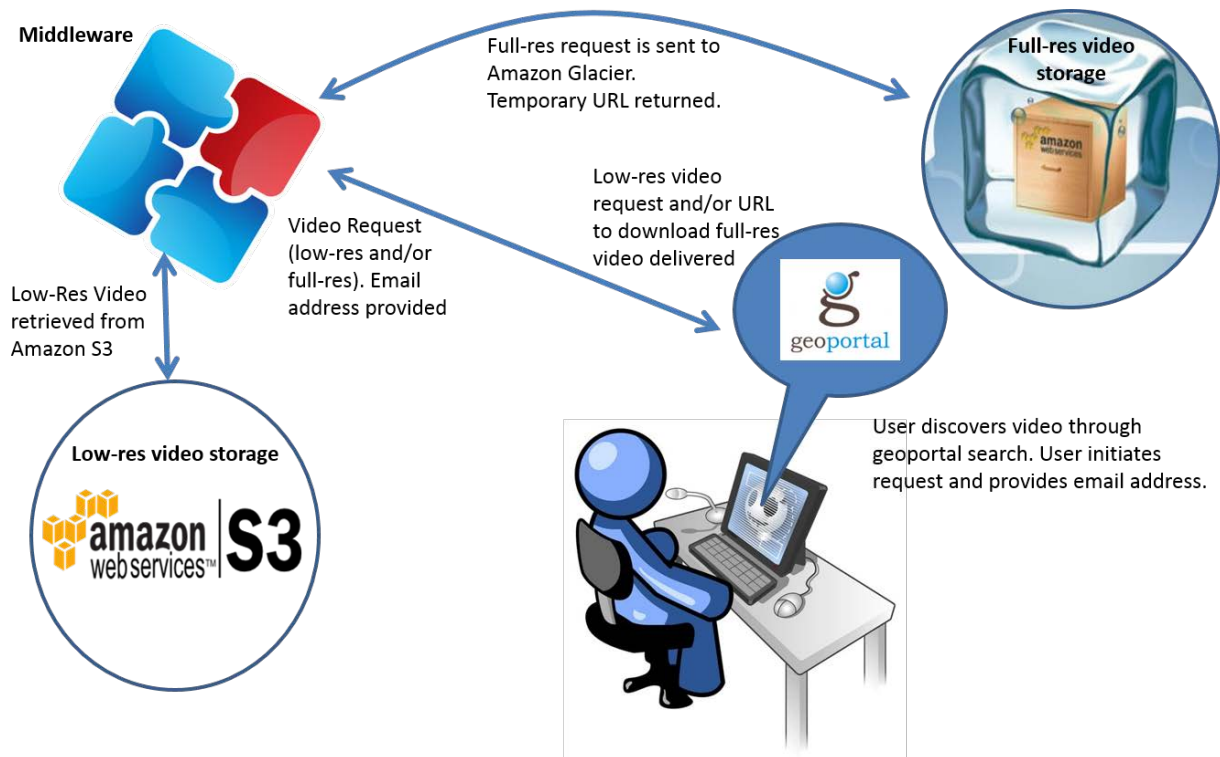
Figure 1: The OER Video Cloud Pilot Project schematic

IV.   VDMMI Project Implementation (2014-2016)

As a result of the successful pilot, the OER Data Management Team, NCEI, and the NOAA Central Library undertook the Video Data Management Modernization Initiative (VDMMI) project to design and implement a similar, yet customized, capability using NOAA's available deep storage infrastructure, Comprehensive Large-Array Storage System (CLASS), and the persistent online storage available where the lower-resolution video from *Okeanos Explorer* already was stored.

The VDMMI project began with a full inventory of OER video assets. These video assets included those stored on physical media (tape and optical disks) at the NOAA Central Library as well as file-based digital video stored on spinning disks at the Inner Space Center in Rhode Island, in persistent

storage at the NOAA Central Library, on portable hard-drives maintained by the OER Data Management Team, and on the *Okeanos Explorer* itself.

A.   Inventory of Video Data on Physical Media

Video data on physical media presents a particularly challenging set of issues and OER is dealing with a large backlog of video in this form. Over 4,400 physical objects containing video data have been collected since the first OER-funded cruise in 1999. Although these comprise a diversity of formats, the principle formats are DVD (46%) and MiniDV (38%). Altogether they represent over 34 TB of data, yet only a small proportion (2.8%) has been transferred to digital files for preservation and access. Moreover, these have been only in the form of highlight videos and segments, so no full-length 1-1 copies had been created. Transfer is an expensive process, both in terms of labor costs and in terms of storage space. As these media have aged, they have deteriorated, making it more difficult or impossible to transfer to file-based copies. Finally, the current version of the primary tool used to effect tape transfers, Final Cut Pro, has lost several important transfer capabilities as it has been updated, such as logging time code errors and reducing the success rate of tape-to-file transfers. Thus, OER physical media is in a precarious state, compromising NCEI's role to steward and preserve the video data from these legacy expeditions. OER is investigating an outsourcing solution to recovering the video data from physical tape.

B.   Inventory of *Okeanos Explorer* Video Data

OER began capturing underwater video directly to spinning disk during the *Okeanos Explorer* 2010 field season. Spinning disk video is not subject to the same degradation issues as legacy media, but it carries with it its own set of challenges. While eliminating the time and cost bottleneck of real-time transfer to computer files, a combination of technological advances: the use of full video resolutions and low-compression codecs; longer dive times due to the use of unmanned ROVs; multiple simultaneous sources and streams of video (including Telepresence, a high-speed real-time ship-to-shore communication network for directing and sharing video capture) has greatly increased the sheer volume of video acquired on a given dive. This has put enormous pressure on OER to develop capacity to store, edit, manage, and make accessible digital video. Pre-*Okeanos*, one 3-hour dive might have generated 40 GB of video; *Okeanos Explorer* video for an 8-hour dive typically generates over half a terabyte of data. This includes full length video streams, short segments of

interesting subjects determined during the dive, and compilation "dive trailers" comprised of dive video highlights.   By the end of 2016, spinning-disk-based digital video captured during 284 dives launched from 20 ROV cruises on the *Okeanos Explorer,* took up almost 120 TB of space and accounted for approximately 75% of all data managed by OER. Further growth is expected to result in a doubling of this volume by the end of the next field season.

C.   NOAA Video Archive Infrastructure

NCEI selected CLASS, a recognized NCEI archive, to preserve OER's video data for this project. This dataset includes the full-resolution segments, the low-resolution segments, the streamed footage (the video that is streamed to the Internet during a dive from each of the two vehicles), and the highlight videos in low-, high-, and full-resolution. NCEI wanted to use CLASS instead of the Oceanographic Archive System (OAS) for the video and video products that are not in full-resolution because of the volumes of the data and the fact that the volume of these data would unduly stress the required backup procedures. The Common Submission pathway was used to load the video data into CLASS.

During the required CLASS Engineering Assessment, it was recommended that a 5GB file size was preferred and that no file should exceed 8 GB.   Going forward, standard operating procedures were put into place to ensure that no video segment exceeded the 8 GB upper limit. Looking back, however, the team identified all of the segments that exceeded 8 GB and made a plan to programmatically divide the files into 5 GB "chunks" - renaming the segmented segments in the process. Both the low-res version and the full-res version had to be "chunked" so as to keep the one-to-one correspondence required for the system to work. These renamed files also needed to be copied to all of the locations where these video data might reside - on the ship, in the NOAA Central Library, and at the URI ISC - before they were uploaded to CLASS.

Since OER's inception, video data has been preserved at the NOAA Central Library (NCL). NCL had always been a division within the National Oceanographic Data Center (NODC). During the reorganization of NOAA's data centers into the NCEI, the NCL was relocated into the NOAA Oceanic and Atmospheric Research (OAR) line office. As a result, the video data (plus images and event logs) at NCL were migrated to NCEI's spinning disk storage. NCL remains the repository for OER's documentation and reports, which continue to be published in an Institutional Repository.

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 11

Additionally, NCL will update their existing cruise catalogs to point to the new storage locations for the low-res segments, images, and event logs rather than remove them completely.

D.   Machine Readable Formats

Alternate data storage formats were explored by the OER Data Management Team during this project. One such format evaluated was the Hierarchical Data Format Version 5 (HDF5) which was found to be inadequate for OER video data management needs, but is described here as it may be a useful format for archival and delivery of video data for others.

HDF5 is a data model and file format designed to store and manage high volume, complex data and is designed to facilitate the association of data objects with rich, descriptive metadata within the HDF5 file or across multiple HDF5s. The OER Data Management Team determined that HDF5 meets the five criteria set forth by the United States National Archives as an appropriate data format for archival, despite not being on any discoverable list of approved archive formats at the time the research was conducted. The team discovered, however, that there was no technical advantage to using HDF5 as a storage medium for video because a HDF5 video file cannot be used natively in a video player. For more information about the team's HDF5 exploration, refer to the evaluation documentation located [here](#).

E.   Metadata and Geoportals

The undeniable key to success in discovering video is detailed metadata, coupled with a metadata search portal that can accept filtering criteria from the user, search the metadata catalogs and return results. Each cruise on which video data were recorded will have a "collection-level" metadata record published in [NCEI's geoportal](#).   NCEI also maintains a "granular" level geoportal which contains metadata for individual data files – in the case of OER video, the video segments, the video streams, and the video highlights.

The cruise metadata record, and the subsequent NCEI "landing page" for the cruise accession, provides information about the cruise and links to data and information resources. The granular video metadata records, and the subsequent NCEI landing pages for the video resources, provide the name and abstract for each video segment and links to a preview version of the video as well as a

link to place an order for the corresponding full-resolution version. But in order to get to these landing pages, the user needs to be able to discover them – and that is done through the metadata.

A key technique for OER video data management is to embed critical metadata in the filename of each asset. This insures that files will always be understandable and discoverable even if the link to metadata files are lost. This requires a commitment to a standardized naming convention, and makes for some long file names. Using the naming convention OER created for *Okeanos Explorer* video, it is possible to identify from the file name:

- the cruise during which the video was taken, expressed as expedition id (two digit ship code + two digit field season year + two digit sequential cruise number within field season, and optionally an 'L' and a two digit sequential leg number)
- the dive number (for the Okeanos Explorer, this is not a unique number. The Okeanos uses DIVE01, DIVE02, etc for each cruise)
- an indication that the asset is a video file
- the time stamp associated with the first frame of the video, expressed as year, month, day, time in hours, minutes, and seconds, relative to UTC
- the camera and platform from which the video was taken
- codes and free field text describing the content of the video
- the resolution of the video (low, mid, or high, if this attribute is missing, the file is assumed to be full-res)
- the file format container (usually .mov, which indicates a QuickTime wrapper, may be .mp4 in the future).

For example, the video file captured during cruise EX1402, Leg 3, on April 12, 2014 at 19:10:15 (UTC) using the High Definition camera on the main ROV submersible, showing a mound and a shrimp would be named EX1402L3_VID_20140412T191051Z_ROVHD_MOUND_SHI_Low.mov. The "Low.mov" suffix indicates that the video file is a low-resolution QuickTime-wrapped .mov file. The corresponding full resolution video segment would be named identically, but without the "_Low" – EX1402L3_VID_20140412T191051Z_ROVHD_MOUND_SHI.mov.

Another stage of metadata generation requires some information about the cruise on which the ROV operations are occurring. The OER Data Management Team uses an Access database called a Cruise Information Management System (CIMS) to capture much of the metadata needed. Metadata is gathered about the cruise at the planning stage and saved in CIMS. The CIMS then outputs an ISO collection-level metadata record with the "who, what, when, where, why, and how" of the

cruise. Once the cruise is completed, the OER Data Management Team enters data metrics for each data type collected, including the video. CIMS will export certain header type information in an xml format which is then used by video annotation extraction routines and metadata generation routines. The following list of sources of metadata for each video segment include:

- Cruise header information (cruise title, abstract, dates, principal investigator, ocean basin, ocean sub-basin, geospatial boundaries, mission themes)
- Dive observation notes (dive purpose, dive observation, geospatial boundaries, chief scientist, dive site name, maximum depth)
- Embedded video metadata (duration, wrapper, codec, bitrate, frame rate)
- Video file name (UTC date/time stamp, vehicle, camera, taxa codes, free form text)
- Submersible CTD (establish range values for depth, temperature, and salinity using start time of video segment and duration)
- Submersible Navigation (latitude, longitude, altitude, attitude, pitch, roll, heave)
- Scientific Chat Log (video annotation by participating scientists using code tables and free form text)
- If the video segment captures a specimen collection, the EX Sampling Operations Database Application (SODA) (genus/species, size, condition, weight, collector, comments)
- If the video segment goes viral on social media, the names (in all conceived spellings, of the viral monikers)
- If the video segment contains a frame grab used in the Okeanos Benthic Animal Guide, the Group, Subgroup, Category, Subcategory, Family, Taxonomic ID, Phylum, Subphylum, Superclass, Class, Subclass, Infraclass, Superorder, Order, SubOrder, Infraorder, Superfamily, Subfamily, and Subgenus of the imaged biological individual.

A resource the team relies heavily on is NCEI's Docucomp utility. Docucomp is an ISO metadata component library that allows account holders to build metadata snippets that tend to be relatively static. Universally Unique Identifiers (UUIDs) can be generated through online UUID generators and then included in the metadata snippet before saving. Such snippets include CI_ResponsibleParty metadata segments for organizations and for personnel; MI_Instrument metadata segments for instruments; MI_Platform for vessels and submersibles; and MD_Keywords for standard vocabularies. These are then referenced in the metadata records as calls to the component library. These "unresolved" records are then resolved through the Docucomp before being saved and published. This will take a snapshot in time of the details for each of these metadata snippets. If information changes for any of them, the updates are made in Docucomp and future metadata

records will get the new information. If older metadata requires the new information, re-resolving will perform that function.

All of these data are then mapped into the ISO 19115-2 formatted metadata template that the team designed and refined over the course of the project.

F.   OER Video Portal

The OER Video Portal is a user-friendly online tool for search, discovery, and access of all OER video. The user interface front page presents a search engine tool with multiple fields for identifying desired criteria for video data. Given a position of prominence at the top left of the page is the search field for keywords, perhaps the most critical field for browsing the collection. Keywords have been extracted from event logs; cruise documents; transcriptions of online conversations with ship personnel, researchers, and telepresence-enabled shoreside collaborators; and codes in the file names, which follow strict conventions for pre-defined taxa. Users can choose keywords either singly or in groups from an auto-completed drop-down list, or by free text entry. Various options allow users to drill down to preferred levels of specificity, for instance, keywords can be concatenated or matches can be required to be exact.

In addition to keyword search, there are fields for
- temporal data, representing the start and end date for desired video;
- geospatial coordinate boundaries through use of a map or by defining the boundaries;
- minimum and maximum depth in meters of the vehicle from which the video is being recorded;
- location areas, defined by the science planning teams, such as "Bryant Canyon Deep";
- cruise names, represented in both English descriptors ("ROV Exploration of Atlantic Canyons and Seamounts") and as Expedition Year and Leg (EX1504L3).

Both locations and cruises are selected from drop-down menus. Even casual users might be aware of these terms from browsing videos on the *Okeanos Explorer's* own web page,.

Once the user is satisfied with the level of search values and submits the search, the OER geoportal processes the request, returning a list of candidate video segment results that can be individually or collectively selected. Granule-level ISO metadata provides the source from which several helpful

attributes are displayed: title, abstract, key frame thumbnail, and links to the video data landing page (video details), a dive summary report (if available) and to a video preview pop-up window from which the low-resolution video can be streamed for the user.

If the user wants to be able to download or order videos using the portal, the individual video segments can be checked to place them in the user's "basket." Additionally, if the user wants to place all of the videos in the results in their basket, there is a checkbox to put them in 100 at a time. The maximum number of files that can be ordered in one order is 1000 segments.

After selection, the user can "view basket" to see all of the videos ready for ordering or downloading. Users must "order" full-resolution videos but the low-resolution videos can be downloaded in bulk into a zip file. Full-res files, since they are stored nearline, can be ordered by providing an email address. Depending upon volume, but usually within a few minutes, users receive an email directing them to an ftp site from where they can download the requested files. See Figure 2 for a schematic.
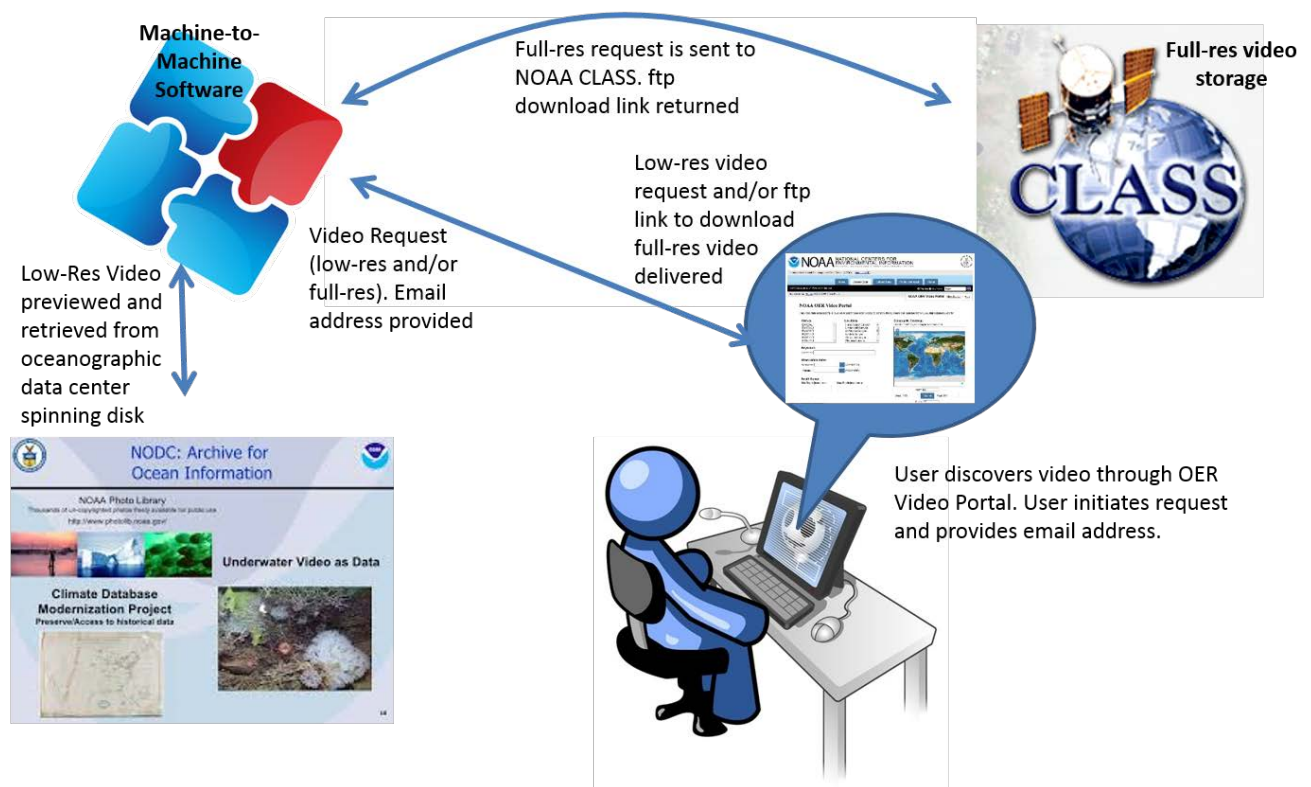

Figure 2: OER Video Data Management Modernization Initiative schematic

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

V.   Conclusion and Best Practices

While the VDMMI project has provided a much-needed solution to self-service discovery of and access to *Okeanos Explorer* video data assets, challenges remain. In addition to the need to rescue legacy data, a secondary and equally important effort will address integration of digital video from non-Okeanos system collections, where annotation may be less specific.

Programs with video data assets looking for a similar solution are encouraged to adopt methods described and materials offered through this effort. The OER Data Management Team also recommends several best practices that will help ensure a successful transition to a capable, extensible, technology-leveraged video data management system.

- *Open access to scientific data for all is a target state that should drive development*
    - The OER video data collection represents a high volume, high value data set, comprising a wide product range and a diverse and wide user base with different requirements. To maximize the value of this dataset, all users should have open access to this data.
    - The 2015 NOAA plan to increase Public Access to Research Results (PARR), written in response to the White House Office of Science and Technology Policy (OSTP) Memorandum of February 2013, requires public accessibility to all data and publications produced by federal researchers or by recipients of federal funds, such as the OER video data collection.
    - Open access is leveraged by advances in technology, evolution of metadata standards, and social uses of scientific data; therefore, it will be necessary to continually revisit development of the tools created to achieve it.
- *The widespread availability of rich metadata is essential for a successful, unified method of managing a large video data collection*
    - Reliable, complete geospatial, temporal, environmental and keyword-based metadata are necessary for self-service discovery and access
    - Metadata must be gathered starting early in the planning stages and continuing through the entire lifecycle of video asset use
    - Metadata should be associated with both collection-level and granule-level structures in a hierarchical model that allows users to traverse easily between the two
- *Video should be captured at the highest quality levels with the lowest level of compression consistent with available resources and expected downstream processing and use*

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 17

- o *Okeanos Explorer* video is captured at 1920 x 1080i employing the ProRes 422 video codec inside a QuickTime container format, at a bit rate of 145 Mbps. This format utilizes mild compression but is still considered broadcast quality.
  - o SMPTE Timecode referenced to UTC should be embedded in the video signal, fed from an aboard-ship master clock to synchronize all video sources during dives.
  - o SMPTE Timestamp should be made part of the video filename to aid discovery and to enable metadata extraction from other sources
- *Video storage and preservation requires a large capacity infrastructure that can handle current and future needs*
  - o Move video files to stable storage media as soon as possible to avoid deterioration
  - o Long term storage and preservation of *Okeanos Explorer* video is housed in the NOAA CLASS infrastructure, which has enough capacity to meet current and expected demand for the scientific and broadcast communities.
  - o CLASS data is backed up frequently to insure long term preservation of archived video.
  - o To support discoverability, accessibility, and usability, online streaming or downloading of preview copies is served from spinning disk. These copies are also stored in CLASS for long-term preservation and backup.
  - o A well-designed middleware geoportal should allow for immediate preview of online video and short-term ordering of near-line video.
- *Standards-based approaches ensure interoperability*
  - o Metadata conforms to ISO 19115-2 standard for image-based geospatial data. This facilitates the hierarchical linking of each granule-level video segment through its metadata record to the parent collection which is the cruise during which the video was captured.
  - o Each cruise is assigned a stable, persistent Digital Object Identifier (DOI), minted by the California Digital Library's EZID system. Assigning a DOI to OER video ensures that it can always be discoverable, even if the URL linked to it changes for any reason.
- *Pilot projects on a small scale before committing to full-scale implementation*
  - o Information learned during the FADGI study gave the VDMMI team the confidence that video capture was robust and that it was time to move onto the storage and dissemination phase.
  - o The Amazon Cloud pilot gave the VDMMI team the confidence that the core concept of two tiers of storage, near-line tape and online spinning disk, linked by a middleware geoportal, was sound. However, it was critical that the storage system

have 100% preservation durability in order to fulfill NOAA's mission of stewardship for the future.

o External presentations of the system in the last few months before rollout were critical to test out the viability of the new system before other oceanographers and data scientists.

o Feedback gained at the June 2016 presentation to the Underwater Video Workshop held at the University of Rhode Island was especially helpful, as the audience was largely comprised of peers holding positions at other underwater video archives, such as the Scripps Institution of Oceanography, Woods Hole Oceanographic Institution, Ocean Networks Canada, NASA Ames Research Center, National Science Foundation, and the Schmidt Ocean Institute.

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 19

Appendix A. FADGI Study

OER's pilot team first joined a Federal Agency Digitization Guidelines Initiative (FADGI) Audio-Visual Workshop Group study to participate in forming the recommendations from industry and government experts on how to best steward video data and then to apply those best practices to the OER video collection.

The FADGI study was a collaborative effort by several federal agencies, including the Library of Congress, the National Archives and Records Administration, and the Smithsonian Institution, to define common guidelines, methods, and best practices for both creators and archivists. OER's underwater video data from *Okeanos Explorer* formed one of the eight use studies.

The cloud project infrastructure constituted a two layer storage model. In alignment with FADGI recommendations, video long term storage would be in the same high resolution format as originally captured, ProRes 422, thus avoiding additional stages of compression and decompression. However, due to file sizes, transmission demands, and lack of user tools for viewing, this data would be put in Amazon Glacier "near-line" tape storage, where some retrieval latency could be tolerated, on the order of 4 hours to two days. Lower-res access video, that is, video of a resolution sufficiently small enough for frequent access and web viewing, would be served from Amazon's S3 persistent, low-latency, on-line disk storage. At the same time, Amazon would automatically shuttle high demand full-res video from Glacier to S3 in order to make it make it available multiple times without having to repeatedly retrieve it. This was expected to save money since there was a high per-use retrieval cost.

FADGI participation was critical to the pilot project because it helped define high-level recommended practices that could be translated into a capture-to-discovery workflow. In particular, it helped put upstream decisions on a sound footing so that downstream processes such as storage, preservation, and dissemination, might be best served. The final FADGI report was issued in December 2014.

References

"An evaluation of Hierarchical Data Format Version 5 (HDF5) as a Storage Model for Archive and Delivery of Video Data" http://oer.hpc.msstate.edu/vdmmi/HDF5_Video.pdf

Comprehensive Large Array-data Stewardship System (CLASS) "Common Submission Interface Control Document" http://service.ncddc.noaa.gov/rdn/oer-waf/media/docs/1469 CLASS Common Submission Interface Control Document FINAL v1.2.pdf

Docucomp: The home of component metadata authoring http://www.ngdc.noaa.gov/docucomp

Federal Agencies Digitization Guidelines Initiative: Creating and Archiving Born Digital Video http://www.digitizationguidelines.gov/guidelines/video_bornDigital.html

NCEI Ocean Archive Geoportal
https://data.nodc.noaa.gov/geoportal/catalog/search/search.page

The OER Digital Atlas
http://explore.noaa.gov/digitalatlas

The OER Data Management Team's Metadata Template
http://service.ncddc.noaa.gov/rdn/oer-waf/media/templates/EX_Video_Segment_Template.xml

The OER Video Portal
http://www.nodc.noaa.gov/oer/video/

The *Okeanos Explorer*'s webpage
http://oceanexplorer.noaa.gov/okeanos/explorations/explorations.htm

The Report of the President's Panel for Ocean Exploration "Discovering Earth's Final Frontier: A U.S. Strategy for Ocean Exploration" http://oceanexplorer.noaa.gov/about/what-we-do/work-areas/ocean-panel-report.pdf

Acknowledgements

Points of Contact

Dr. Alan Leonardi, Director
NOAA's Office of Ocean Exploration and Research (OER)
Alan.Leonardi@noaa.gov

Jeff de La Beaujardiere, PhD
NOAA Data Management Architect
Jeff.Delabeaujardiere@noaa.gov

Sharon Mesick, Chief, Information Services
NOAA's National Centers for Environmental Information
Center for Coasts, Oceans and Geophysics
Sharon.Mesick@noaa.gov

Susan Gottfried, OER Data Management Coordinator
General Dynamics Information Technology
Supporting NCEI and OER
Susan.Gottfried@noaa.gov

Brendan Reser, Video Data Management Project Technical Lead
Scientific Technology, Inc
Supporting NCEI and OER
Brendan.Reser@noaa.gov

OER Data Management Team Report
Video Data Management Modernization Initiative (VDMMI)
December 2016
NOAA Ocean Exploration and Research Data Management Team (oer.info.mgmt@noaa.gov)

Page 22