NOAA Technical Memorandum NMFS-AFSC-514

Curation of Alaska Marine Mammal Observer Program (AMMOP) Legacy Data to Improve Access and Useability

M. Moon and R. Burns

March 2026

U.S. DEPARTMENT OF COMMERCE

National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Alaska Fisheries Science Center

The National Marine Fisheries Service's Alaska Fisheries Science Center uses the NOAA Technical Memorandum series to issue informal scientific and technical publications when complete formal review and editorial processing are not appropriate or feasible. Documents within this series reflect sound professional work and may be referenced in the formal scientific and technical literature.

The NMFS-AFSC Technical Memorandum series of the Alaska Fisheries Science Center continues the NMFS-F/NWC series established in 1970 by the Northwest Fisheries Center. The NMFS-NWFSC series is currently used by the Northwest Fisheries Science Center.

This document should be cited as follows:

Moon, M., and Burns, R. 2026. Curation of Alaska Marine Mammal Observer Program (AMMOP) legacy data to improve access and useability. U.S. Department of Commerce, NOAA Technical Memorandum NMFS-AFSC-514, 18 p.

This document is available online at:

Document available: https://repository.library.noaa.gov

Curation of Alaska Marine Mammal Observer Program (AMMOP) Legacy Data to Improve Access and Useability

M. Moon[1] and R. Burns[2]

[1]Alaska Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
7600 Sand Point Way NE
Seattle WA 98115

[2] Pacific States Marine Fisheries Commission
6720 S Macadam Ave, Suite 200
Portland, OR 97219

U.S. DEPARTMENT OF COMMERCE

National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Alaska Fisheries Science Center

NOAA Technical Memorandum NMFS-TM-AFSC-514

March 2026

ABSTRACT

The Alaska Fisheries Science Center's (AFSC) Fisheries Monitoring and Analysis (FMA) Division is responsible for training, in-season support, and quality control of observers who collect catch data on board fishing vessels and at shoreside processing plants. The FMA Division undertook a comprehensive project to curate and manage legacy data associated with the Alaska Marine Mammal Observer Program (AMMOP). These data documented marine mammal and bird interactions, as well as catch associated with the commercial salmon gillnet Alaskan fisheries. The work involved data standardization, quality assurance, and migration to a modern infrastructure. The work increased the integrity, accessibility, and utility of the AMMOP data, and preserved the historical record for contemporary scientific research.

The FMA Division collected the AMMOP legacy information from several sources, determined uniqueness, and inventoried the files and folders. Fishery data were organized into relational databases. We documented the inventory process and database changes in a curation log. We created validations and cleaned the data for use in analyses. Project completion created a consolidated, reliable, and easily searchable data repository, enhancing the ability of the NOAA scientists and managers to use and add to this foundational observational data of marine mammal and bird interactions with fisheries in Alaska. The AMMOP curated data and supporting documents are available through the FMA Division's confidential data request process. Please contact the lead author for more details.

CONTENTS

ABSTRACT iii

INTRODUCTION 1

METHODS 1

Inventory 1

Pooling 1

Preparing files 2

Determining uniqueness 2

Logging 2

Databases 3

The AKFISH_PRD and AMMOP schemas 3

Validation 3

RESULTS AND DISCUSSION 4

ACKNOWLEDGMENTS 7

CITATIONS 9

APPENDICES 17

INTRODUCTION

The Alaska Marine Mammal Observer Program (AMMOP) was deployed in ten different years between 1990 and 2013 at 6 geographic locations (Table 1). The AMMOP deployments used observers (biologists) to collect data on marine mammal and bird interactions, as well as catch associated with the commercial salmon gillnet Alaskan fisheries (Dietrich et al. 2025). Gillnets in the Alaska salmon fisheries were set either by drifting the net from a small vessel (driftnet) or stretching the gillnet from shore and using skiffs to tend the net (setnet). Legacy information for the 1990 Unimak and 1991 Prince William Sound deployments consists only of summary reports. Legacy data for the 1999 and 2000 Cook Inlet deployments were entered into Microsoft Access™ databases by the observers during the deployments. Deployments for the 2002 and 2005 Kodiak, the 2007 and 2008 Yakutat and the 2012 and 2013 Southeast Alaska were stored in an Oracle® schema containing two relational databases.

The Alaska Fisheries Science Center's (AFSC) Fisheries Monitoring and Analysis (FMA) Division is responsible for training, in-season support, and quality control of observers who collect catch data on board fishing vessels and at shoreside processing plants. In 2021 the Fisheries Monitoring and Analysis (FMA) Division received the AMMOP existing Oracle schema from the Alaska Regional Office (AKRO). The FMA Division was selected to maintain the AMMOP legacy data as it had conducted the North Pacific Observer Program for decades, as well as housed, maintained and provided fishery data to analysts for years. Additionally, the FMA Division was planning future deployments of the AMMOP.

Data requests for the AMMOP legacy data received by the FMA Division demonstrated the difficulty of finding and interpreting the data for specific requests. The supporting documentation for the AMMOP was varied in format and sometimes missing. In 2024, the FMA Division was awarded a grant from the National Oceanic and Atmospheric Administration (NOAA) Fisheries Information System program to curate the AMMOP legacy data and supporting information.

We conducted a curation project where we organized and documented all findable information for the AMMOP, reviewed and updated database designs within a new Oracle schema, and validated and logged exceptions found in the data. We ensured our work followed the NOAA mandates that all of its data are "managed following the FAIR principles to ensure it is Findable, Accessible, Interoperable, and Reusable…" (NOAA 2023). The goal of FAIR is to enable others to use and build upon existing data (Wilkinson et al. 2016). Work began in April 2024 and completed in March 2025.

METHODS

Inventory

For clarity in the inventory methods below, we refer to all information received as files.

Pooling

We requested files from individuals known or believed to be points of contact for the AMMOP legacy information. We called these points of contacts stewards. The files obtained from stewards were stored in a secure Google Drive™ folder. When stewards shared their files of legacy data with us by Google Drive, we requested an editor role and moved the files to our secure folder; in cases when the steward wanted to maintain their files on the Google Drive, we requested they remove our access once we had copied their files (Fig. 1, Flowchart color legends and symbols are in Appendix 1 and Appendix 2,

respectively.). This workflow created a check, in which, if we discovered AMMOP files on the Google Drive outside of our curation folder, we considered the files as new and added them to the inventory process. We organized all files by steward in a folder labeled 1_file_drop_off. A backup of the pooled files was created in the subfolder of original_data_do_not_alter (Table 2).

Preparing files

Microsoft Windows (windows) was used for organizing and housing the inventory files because it streamlined file movement, provided easier understanding of organization and inventory status, and reduced the third-party priority software needed to house the files in the future. All files in the Google Drive folder were downloaded into our secure windows folder. All zipped folders were extracted, and the original zipped folders were moved into the redundant_inventory folder (Fig. 2). We created an organized-stepped folder structure that followed the workflow of sorting and isolating files; this structure allowed a quick understanding of files received and each file's status in the inventory process (Table 2, Fig. 2). Additionally, this structure minimized the number of times a file had to be reviewed during the inventory process.

Within folder 2_pre_inventory we moved and organized the files from 1_file_drop_off into subfolders for each the AMMOP deployments (Fig. 3). Within each deployment, files were further organized by steward. Files that did not fit within a deployment were placed in a newly created folder called additional_materials.

Determining uniqueness

To isolate unique records from the pooled legacy data, the team developed R scripts to compare files across different source folders (Table 3). The process involved iteratively comparing each file source against the others. The accuracy of these scripts were verified by running each process on unique and duplicate files we created prior to running on the pooled legacy data (Fig. 4).

For the initial script we organized files into three different folders. These folders were Original, for the deployment files from a single source, Control which was a copy of the Original folder, and Treatment which was the deployment files from a different source. The initial script, unique_file_working.R, compared a widely used file identifier known as an MD5 hash, between the contents in the Original and Treatment folders to detect duplicates. We also compared the Original and Control folders to detect issues with folder set up and code. The script exported three .csv files, which were saved for supporting documentation. Duplicate files were moved to the redundant_inventory folder. To capture missed files, we ran the script hash_count_check.R between remaining unique files between folders and exported the results for supporting documentation. Microsoft Excel™ files determined to be unique from the first 2 scripts, but that we suspected were duplicates, were further processed using script xl2xl_compare.R with the R package waldo. Files that were found to be duplicate data were also compared visually as waldo did not compare graphs or cell functions.

Logging

Throughout the process of sorting and isolating the unique set of information, we documented the files and inventory steps in an inventory log. Within this log we captured file source, date, location, final folder, duplication justifications, and other relevant information. Every file in 3_inventory_in_progress was reviewed and described in the log. The organized files were moved to their corresponding deployment folder in 4_inventory_complete; this is the folder designated to hold the completed inventory and to use as a reference to clarify future AMMOP legacy questions.

Databases

The AKFISH_PRD and AMMOP schemas

The data the FMA Division inherited from the AKRO was in an Oracle schema labeled AKFISH_PRD. The schema was moved to the FMA Division's Oracle server. The schema housed two databases, one for Kodiak and Yakutat and another for Southeast Alaska deployments. Some tables were shared by both databases, while others were specific to Kodiak and Yakutat, or Southeast Alaska. The AKFISH_PRD schema also contained non-observer created summary tables of the data; these summary tables were not curated since the information could be obtained from the fishery data tables.

In preparation for curation, two Cook Inlet databases were added to the AKFISH_PRD schema, one for 1999 data and another for 2000. All Cook Inlet fishery data tables were exported out of Microsoft Access as spreadsheets and imported into the AKFISH_PRD schema as Oracle tables. In 2000, Cook Inlet also collected additional information on openers, a period of active fishing for a fishery (Manly 2003). Tables for the openers were added to AKFISH_PRD schema to review for useful information.

All data in the AKFISH_PRD schema was considered pre-curated. Within the FMA Oracle server a new schema, AMMOP, was created for the AMMOP curated databases. A junction table was created within the AMMOP schema to link parent table data back to the corresponding parent table records in the AKFISH_PRD schema. More details on the AMMOP schema are listed by deployment in the next section.

Validation

Validations are checks to ensure data integrity and accuracy. Ideally validations are done during data entry into the database, and periodically as the data is updated. As the data was already entered into our databases, our validations were done to flag potential issues with the data. Although validations were done on curated data, we spot checked the initial source to ensure accuracy and resolve the validation if possible. We used numerous sources to create and focus the validations in the AMMOP schema. These sources included exceptions created and logged into the AKFISH_PRD schema table by prior stewards, feedback from data requestors' analyses of the pre-curated data, prior exploratory queries, and issues found during database creation. Validation types include the following:

- Chronological – Flagged data with dates and times that occurred outside their deployments or outside of other data. For example, if retrieval date and time occurred prior to set, then the record was flagged.
- Duration – Flagged data where duration over 24 hours or less than zero.
- Geographic – Flagged records outside of the study area. Flagged records where net drift speed exceeded 8 knots. Eight knots was selected as this represented an extreme deviation; only 1.84% of the records were over 8 knots.
- Time – Flagged hours and minutes outside of range.
- Missing – Flagged missing data for selected columns.
- Parent-child - Flagged difference between Cook Inlet encounter and haul entered net numbers.

- Species – Flagged Cook Inlet species codes not in the North Pacific Observer Program (NPOP) species list of values tables.

Validations were set up in different packages due to differences between Cook Inlet, Kodiak and Yakutat, and Southeast Alaska databases (Table 3). When the same data check was needed for all deployments, dynamic SQL was used to create processes that adapted to parameter inputs and saved coding time. Utilities needed for the validations were stored in the package AMMOP_UTILITIES. Results and investigation resolutions of validations were stored in the AMMOP_EXCEPTIONS table. The AMMOP schema validations can be refreshed on demand. Validations were organized into four packages within the AMMOP schema:

- AMMOP_VALIDATIONS – Validations that apply to all deployments.
- CI_VALIDATIONS – Validations for Cook Inlet database.
- K_Y_VALIDATIONS – Validations for Kodiak and Yakutat.
- SEAK_VALIDATIONS – Validation for Southeast Alaska.

RESULTS AND DISCUSSION

The FMA Division updated the AMMOP record in the NOAA InPort web documentation repository to include the data steward and curation information. InPort allows users to search and find the NOAA databases and supporting documentation. However, InPort is used within the NOAA and therefore limits access to the AMMOP legacy information which is considered confidential. The AMMOP curated data and supporting documents are available through the FMA Division's confidential data request process. Please contact the lead author for more details.

Inventory

Through our file request process we collected over twenty-six gigabytes that are comprised of over 32,000 files. The files included scans of observer completed field forms and photos of marine mammals and birds. Often the same files were provided by multiple stewards. Our initial haphazard review and organization resulted in unclear file status and documentation, as well as uncertainty of file originality. Our organized stepped-file structure increased our processing speed since files only needed to be handled once, duplicated files were quickly removed, and only unique files were inventoried.

Databases

In the AMMOP schema, we applied the concepts of Third Normal Form (3NF) to fishery databases (Codd 1971). This was done by creating unique rows for data and eliminating redundant information in, and between, tables. When needed data were split into multiple tables to ensure 3NF. Finally, the fishery data were related between tables through primary and foreign keys. We chose to keep the AKFISH_PRD schema and Cook Inlet deployments separated to ensure clarity of datapoint definitions for both the FMA Division staff who pulled the data and those who requested and used the data. Between deployments, the database designs were similar in many cases. However, attempts to combine them would require lengthy metadata and analysts' attention to the variation of definitions of datapoints between deployments. Validations on the AMMOP schema databases flagged many records. In many cases, we were able to use our curation log to locate raw data and resolve the issues.

Some supporting tables were added to the AMMOP schema. These included trip summaries documented by the observers for the 2000 Cook Inlet deployment, the AMMOP to AKFISH_PRD junction table, and species group and research tables from the Kodiak and Yakutat deployments. These tables were provided to allow analysts to determine if the data meets their analytical needs and elucidate unforeseen questions about the data.

Creating a database for the Cook Inlet data took additional time compared to the other deployments. The Microsoft Access databases did not contain many constraints and table relations that would have prevented data entry issues. Lack of proper primary and foreign keys meant when parent value was updated in the parent tables, the child table records could be orphaned or become out of sync with the parent data. A field form for child data needs a datapoint to link back to the parent form. However, a user interface that automatically maintains the same parent primary key until all child tables are entered would have maintained parent-child relationships within the Cook Inlet databases.

The Cook Inlet Microsoft Access databases LOVs were not used as database constraints. For example, an observer could enter the species code for King Salmon (code 222) but then incorrectly enter 'Chum' for the species name. The lack of data entry boundaries created many records with mismatches between the species codes and the species names, as well as typos. We found and corrected records that were entered into Microsoft Access database as the row ID numbers of the table rather than the species code. Requiring a selection from an LOV during data entry would have prevented code and description mismatches, and would have saved time by only requiring the LOV code be entered.

Recommendations

We break recommendations into sections for the curated databases and for potential new deployments. Cost and time are factors. The recommendations are listed by priority.

Legacy data

1. Increase validation creation and resolution. Review scans of raw forms as needed.
2. Continue to update table and column metadata as knowledge improves.
3. Create views of common data requested.

While curating the legacy data, another database and sampling methods were simultaneously developed for an AMMOP deployment in 2025. Due to the loss of resources development was put on pause before completion. We documented the Entity Relationship Diagram (ERD) for the incomplete new database in our inventory. Functional requirements were written from this ERD. Many of the issues found in the legacy data will not occur in the future deployment simply because the new database has constraints, LOVs, and was created following 3NF.

Future AMMOP deployments

1. Ensure new databases are set up following 3NF.
2. Only require the observers to record raw data points. Summaries such as counts and calculations such as soak duration can be calculated automatically from raw data without error.
3. Validate data at the time of entry to prevent bad data.

4. Provide observers feedback validations in near real-time to allow feedback during the fishery.

5. Do not require the observer to collect the same information more than once. For example, date of opener needs only be on the parent paper form and only entered into the database once per unique parent form record.

6. Incorporate independent quality control reviewers to review all data and interview observers. The FMA Division debriefers currently do this for observer collected Magnuson-Stevens commercial fisheries data.

ACKNOWLEDGMENTS

This project was funded through a competitive award by the NOAA Fisheries Information System program, a state-regional-federal partnership supporting sound, science-based fisheries management. Lacey Jeroue provided insight into available data. Martin Park provided the AMMOP future database ERD and information on the design. Kim Dietrich and Kathy Kuletz provided recommendations from their analysis of bird bycatch in AMMOP data. Analysts and former data stewards Brian Belay, Brain Fadely, Brian Lieb, Nancy Young, Shay Howlin and Suzie Teerlink provided data and answered questions to help in our search and understanding. Jason Jannot provided feedback to improve clarity and quality for a draft of this manuscript. Thank you to Andy Kingham, Craig Faunce, and Jim Lee whose reviews greatly improved this document.

CITATIONS

Codd, E.F. 1971. Further Normalization of the Data Base Relational Model. Research Report / RJ / IBM / San Jose, California RJ909.

Dietrich, K.S., Kuletz, K.J. and Moon, M.A. 2025. Marine bird bycatch in Alaska salmon gillnet fisheries. Fisheries Management Ecology. e12809. https://doi.org/10.1111/fme.12809.

(NOAA) National Oceanic and Atmospheric Administration. 2023. Management of NOAA Data and Information. Available: https://www.noaa.gov/organization/administration/nao-212-15-Management-of-NOAA-Data-and-Information.

Manly, B.F. 2003. Incidental Catch and Interactions of Marine Mammals and Birds in the Cook Inlet Salmon Driftnet and Setnet Fisheries, 1999-2000. Western EcoSystems Technology Inc. report

R Core Team. 2025. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

Wickham, H., and Bryan, J. 2025. readxl: Read Excel Files. R package version 1.4.5. https://CRAN.R-project.org/package=readxl.

Wickham, H. 2024. waldo: Find Differences Between R Objects. R package version 0.6.1, https://CRAN.R-project.org/package=waldo.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 160018. https://doi.org/10.1038/sdata.2016.18.

.

Table 1. -- Overview of the AMMOP legacy deployments and data.

Geographic Location	Years On Water	Gear used in deployment	Format data was provided	Curated database object prefix
South Unimak Island	1990	Driftnet; Setnet	Not available; only the final report remains	-
Prince William Sound	1991	Driftnet; Setnet	Not available; only the final report remains	-
Cook Inlet	1999, 2000	Driftnet; Setnet	Microsoft Access	CI_
Kodiak	2002, 2005	Setnet	Oracle AKFISH_PRD schema	K_Y_
Yakutat	2007, 2008	Setnet	Oracle AKFISH_PRD schema	K_Y_
Southeast Alaska	2012, 2013	Driftnet; Setnet	Oracle AKFISH_PRD schema	SEAK_

Table 2. -- Purposes and outputs for folders and schemas used to curate and standardize the AMMOP legacy data.

Object Name	Object Type	Purpose	Primary Result or Product
1_file_drop_off	Folder	To organize all files by source (steward).	Files organized by steward; All found files placed here.
2_pre_inventory	Folder	To move and organize files from 1_file_drop_off into subfolders for each of the AMMOP legacy deployments.	Files organized by deployment and source.
3_inventory_in_progress	Folder	To hold unique files after they have been isolated and to have all files reviewed and described in the log.	Unique files ready for review/logging; Files reviewed and described in the log.
4_inventory_complete	Folder	To hold the completed inventory and to use as a reference to clarify future the AMMOP legacy questions.	Organized files ready for reference.
original_data_do_not_alter	Folder	A copy of 1_file_drop_off as it was prior to the start of file processing. Used to create a backup of the pooled information.	Backup of pooled information.
inventory_process	Folder	The main inventory; contains the organized-stepped folder structure for sorting and isolating files.	Organized-stepped folder structure; minimized file handling.
redundant_inventory	Folder	To hold the original zipped folders after extraction, and to hold duplicate files.	Moved original zipped folders; Moved duplicate files.
additional_materials	Folder	To place files that did not fit within a deployment.	Organized non-deployment specific files.
unique_compare_result_files	Folder	To save the output .csv files from unique_file_working.R for supporting documentation.	Three Microsoft Excel files (control check, mismatch flag, comparison file).
AKFISH_PRD	Schema	To house the pre-curation raw the AMMOP data for Kodiak, Yakutat and Southeast Alaska deployments inherited by the FMA Division. Cook Inlet was initially added to this schema.	Raw and pre-curation data for the AMMOP deployments.
AMMOP	Schema	To house the redesigned databases and validated data.	Validated, curated AMMOP data, with junction table back to the AKFISH_PRD schema.

Table 3. -- R scripts used in detecting duplicate files and validations used to flag database records.

Name	Type	Purpose	Primary Result or Product
unique_file_working.R	R script	The initial script used to compare MD5 hash between files within the original folder to treatment folder to detect duplicates. Also compared original files to a control folder to check for code issues.	Three .csv files saved in unique_compare_result_files: 1. A control check file confirming original and control folders match. 2. A file flagging any original and control files that did not match. 3. A file comparing original and treatment folders, used to locate and remove duplicates.
hash_count_check.R	R script	To confirm all files were unique after the initial check and to catch any files included in multiple deployments.	An exported .csv file of test results.
xl2xl_compare.R	R script	To detect cell-level differences between legacy Microsoft Excel files that were suspected to be duplicates, using the R package waldo. Also used to compare tables exported from Cook Inlet Microsoft Access databases.	Identification of duplicate Microsoft Excel data.
AMMOP_VALIDATIONS	Validation	Validations that apply to all AMMOP deployments across time and space	Flagged records added to AMMOP_EXCEPTIONS database table.
SEAK_VALIDATIONS	Validation	Validations for Southeast Alaska	Flagged records added to AMMOP_EXCEPTIONS database table.
K_Y_VALIDATIONS	Validation	Validations for Kodiak and Yakutat	Flagged records added to AMMOP_EXCEPTIONS database table.
CI_VALIDATIONS	Validation	Validations for Cook Inlet	Flagged records added to AMMOP_EXCEPTIONS database table.

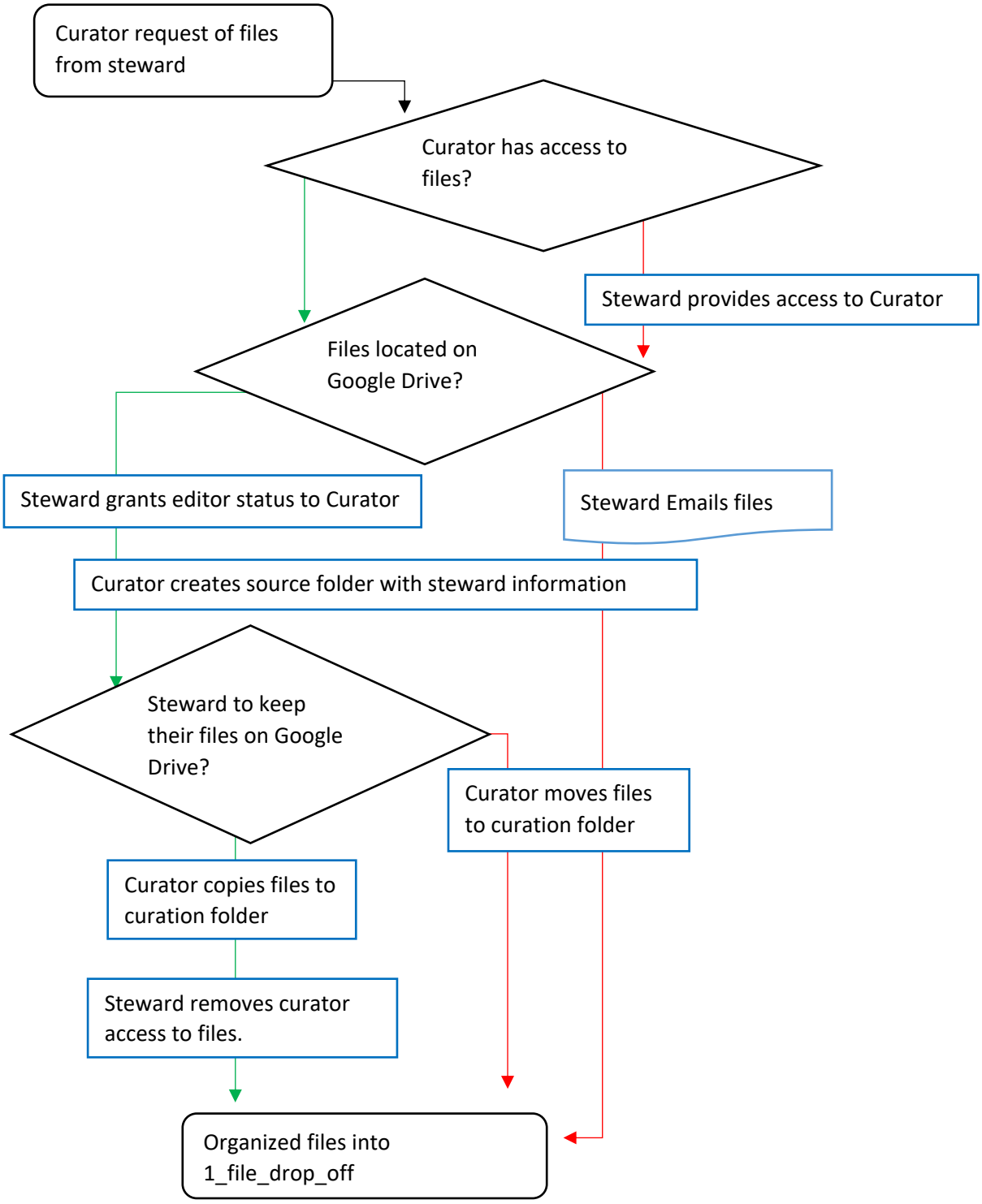


Figure 1. -- The flow of files from request to organized by source in within 1_file_drop_off folder See Appendix 1 and Appendix 2 for flowchart color legend and symbols, respectively.

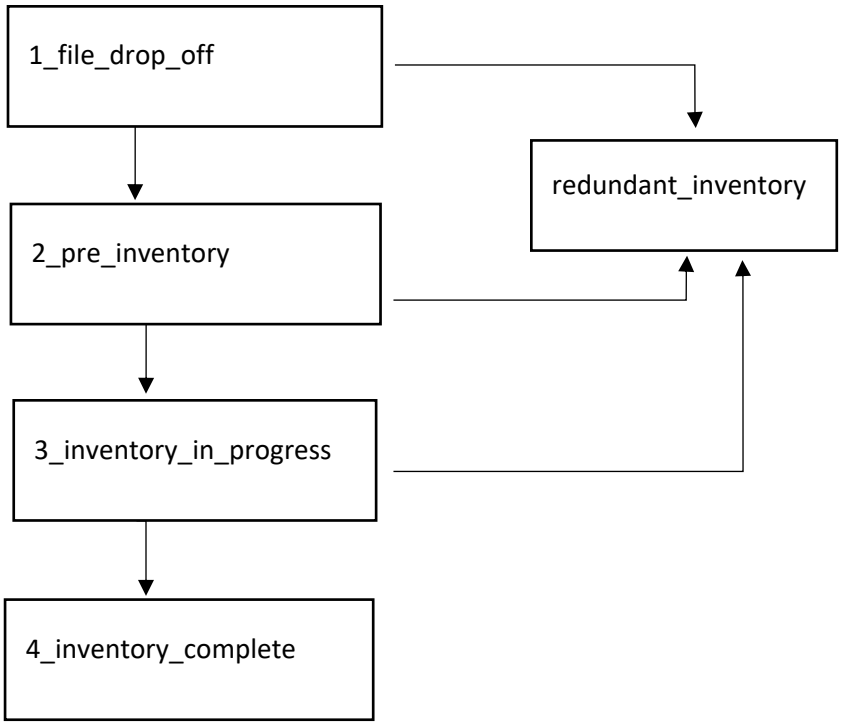


Figure 2. -- Microsoft Windows secure inventory organized-stepped folder structure.

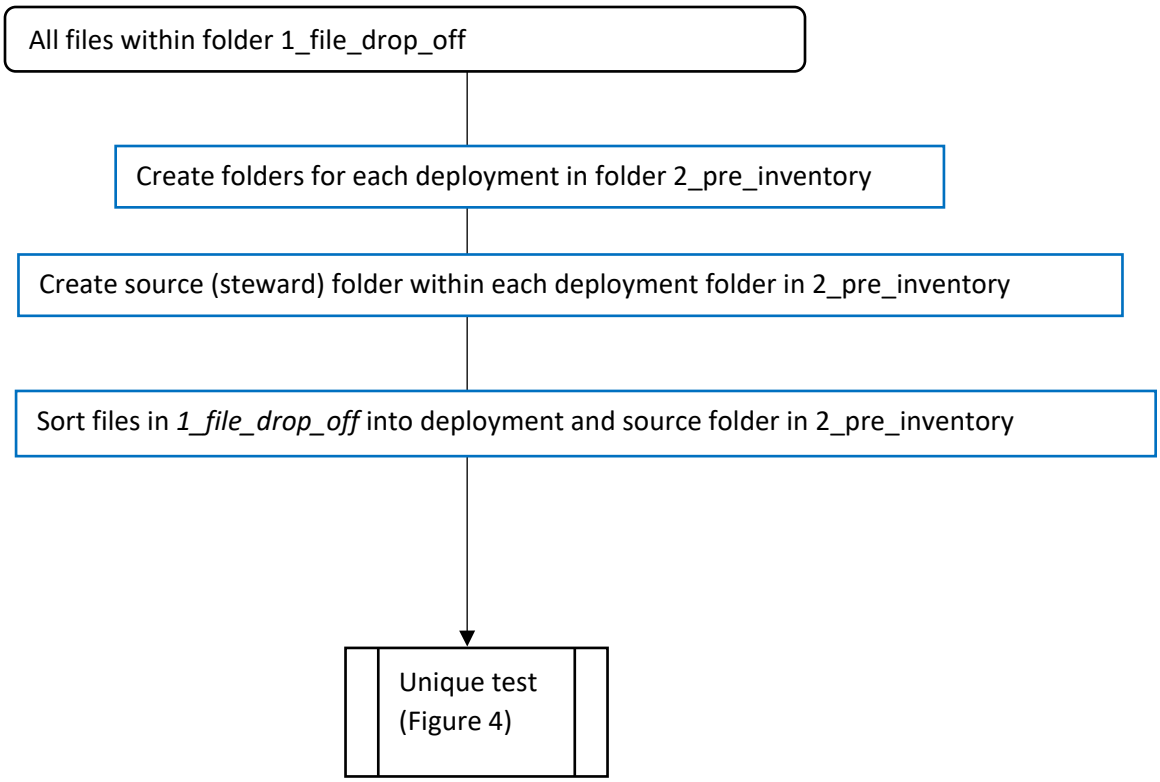


Figure 3. -- Pre-processing subfolders for each of the AMMOP legacy deployments. See Appendix 1 and Appendix 2 for flowchart color legend and symbols, respectively.

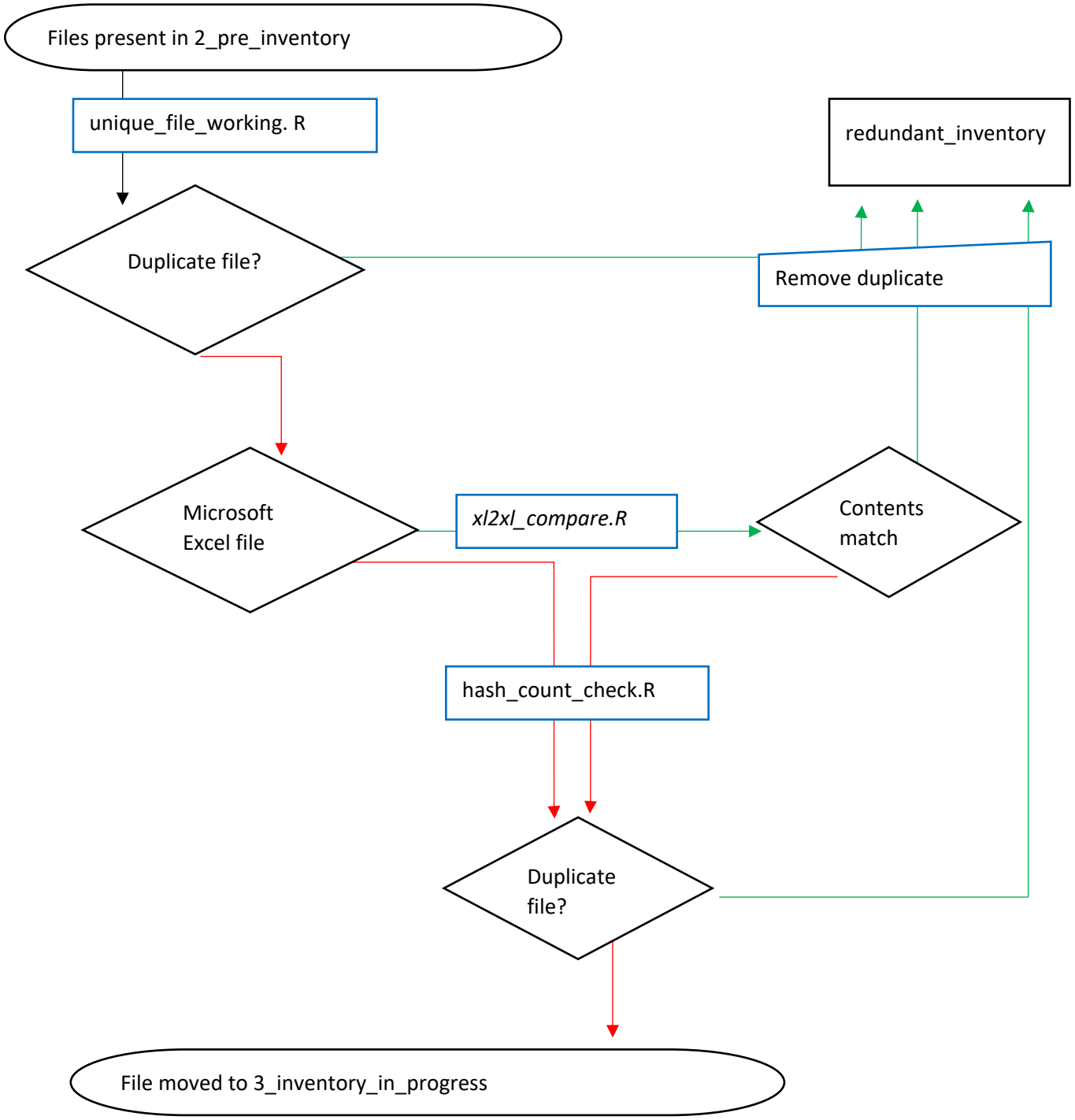
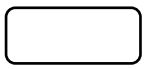


Figure 4. -- Pooled files uniqueness check. See Appendix A and Appendix B for flowchart symbols and color legend, respectively.

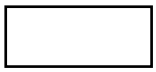
APPENDICES

Appendix Table 1. – Flowchart color legend.

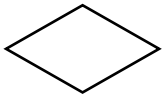
Color	Flowchart
Red	Arrows indicate “No” responses to decisions
Blue	User entered information
Green	Arrows indicate “Yes” responses to decisions
Purple	NA
Black	Automated processes
Grey	Unavailable but present processes



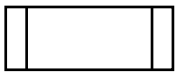
Begin or end of the flowchart diagram



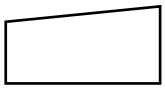
Process



Decision

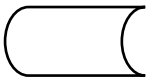


Subroutine or process



Manual input

Data. Usually, process returning data for application or user.



Stored data

Data displayed in application for user to read.

Document (Email)

Appendix Figure 1. – Flowchart symbol legend.

U.S. Secretary of Commerce
Howard Lutnick

Under Secretary of Commerce
for Oceans and Atmosphere
Neil Jacobs

Assistant Administrator, National
Marine Fisheries Service.
Eugenio Piñeiro Soler

March 2026

www.nmfs.noaa.gov

OFFICIAL BUSINESS

National Marine
Fisheries Service
Alaska Fisheries Science Center
7600 Sand Point Way N.E.
Seattle, WA 98115-6349