Chesapeake Bay Program Data Archiving Guidance

Chesapeake Bay Program
Science. Restoration. Partnership.

June 10, 2025

Contents

Introduction

Data resources are foundational to the success of the Chesapeake Bay Program Office (CBPO), with effective data lifecycle management essential to its mission. Creating a framework of defined data lifecycle management processes for CBPO data helps stewards maximize data resource value for stakeholders. The CBPO Data Governance Workgroup has prepared this document to provide clarity and consistency for data stewards regarding data resource **archiving considerations.**

CBPO creates, receives, and publishes several types of data resources. CBPO data stewards follow consistent procedures for data management across its lifecycle, including archiving data resources when content is no longer needed for regular use. Some data resources span several decades, with multiple years of data representing valuable historical content. Other data resources may be superseded when newer versions are created that expand, correct, or update the data. This document was created to assist data stewards in determining when and how to archive data resources (and related products). This guidance directly supports CBPO's Guiding Principles for Data Governance, specifically including discovery and accessibility, classification, quality, value, and transparency. Please refer questions to the CBPO Data Managers.

Purpose

This document is intended to assist CBPO data stewards in determining when and how to archive **authoritative CBPO data resources**. This document is not intended to provide guidance for non-authoritative or working data. A separate document titled CBPO Data Center Strategic Data Plan outlines how operational, non-authoritative data resources should be managed. For more information about that document, please contact the CBPO Data Managers. These guidelines can help data stewards determine which of their authoritative data resources are candidates for long-term data storage, and what that means for accessibility and retrieval times for their content. The criteria can be used as standard principles across all CBPO data stewards.

Data archiving is an important component of the data lifecycle that should be implemented as consistently as other phases of data lifecycle management. Not only does proper data archiving assist users in understanding which data are the most recent version and where those data may be found, but it can also result in significant savings on storage costs when implemented properly. Data stewards may be hesitant to move data to different types of storage if they do not understand the criteria for making those decisions and/or what processes should be followed for placing data into and retrieving it from such storage. As such, the depiction of consistent procedures for archiving data can assist managers in data management, organization, and costs. These guidelines are presented as a universal set of criteria for determining which CBPO data resources are candidates for different types of storage. This document is considered a "living document" that will be updated as needed to support the mission of the CBPO.

Audience

The primary audience for this document includes CBPO data managers, stewards, analysts, and publishers who manage CBPO authoritative data.

Scope

The requirements described in this document pertain to all **CBPO quality-assured authoritative data**, all data that is delivered to the CBPO via grants, and all data that is funded by the CBPO whether direct CBPO funding or indirect (matching funds). Authoritative data may be public or internal; current or legacy; raw or processed. Non-authoritative data, while often crucial in supporting CBPO's mission, are not within-scope of this document. These include data resources obtained from external collaborators or data generated internally that have not undergone a documented QAQC process.

Data Archiving

This section presents key principles regarding data archiving at CBPO, as well as decision points that may be used by data stewards to identify where authoritative data resources should be placed and when they can be considered a candidate for archive. It also presents definitions for key concepts regarding storing and archiving CBPO quality-assured authoritative data.

Key Principles

- **CBPO quality-assured authoritative data** must be retained in **Active Data Storage** for **seven years**. If the data resources and derivative products have not been used in more than seven years, then the content may be moved to the **Cold Data Archive.**
- **Active Data Storage** is used to store **CBPO quality-assured authoritative data resources,** and the contents are instantly accessible to all end users.
- Moving content to the **Cold Data Archive** requires that data stewards review the Cold Data Archive Migration Checklist.
- Content placed in the **Cold Data Archive** must include all related metadata, scripts, and/or models that would be necessary to use and understand the data if the content is restored at a future date.
- **CBPO quality-assured authoritative data resources** moved to **Cold Archive** are never deleted or destroyed.
- CBPO data stewards are responsible for monitoring the age and usage of their data resources, but data resource management and storage activities are the responsibility of the CBPO Data Center.
- CBPO currently implements its **Active Data Storage** environment using Amazon's Simple Storage Service (S3). S3 is a cloud storage service that provides the technology described below for Active Data Storage. Some **CBPO quality-assured authoritative data resources** may be stored outside of CBPO's Amazon S3 Active Data Storage environment; however, the processes and requirements for storing CBPO authoritative data in a robust, instant user access environment and moving that content to Cold Data Archive will remain the same.

General Process Flow

The following steps depict the general process flow when publishing and archiving authoritative data. A visual depiction of these steps is presented in Figure 1 below.

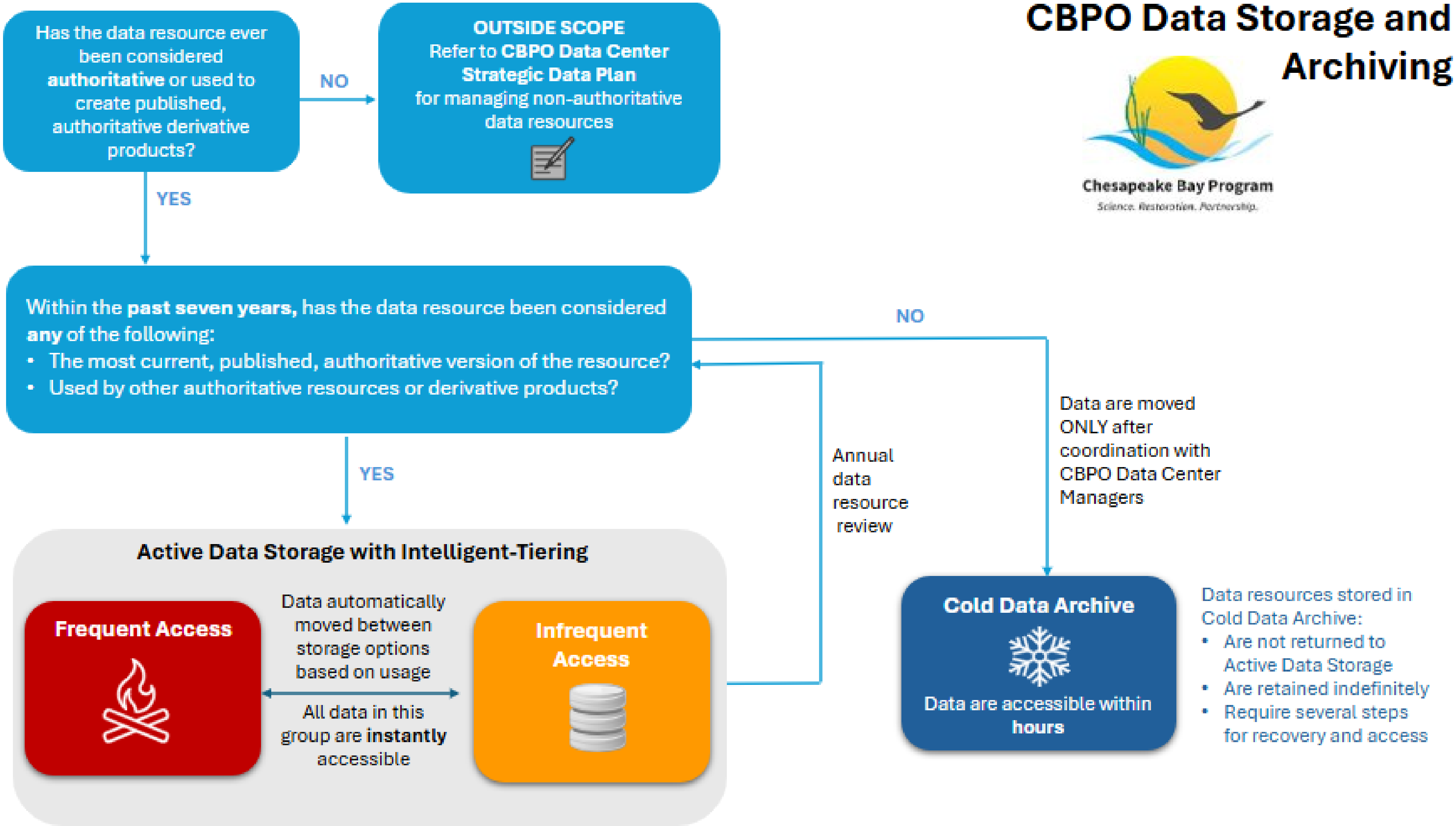
1. If your data resources are considered **CBPO quality-assured authoritative data resources,** then they should be placed into CBPO's **Active Data Storage**

    a. Data will initially be placed into the **Frequent Access Tier** for storage. It will be immediately accessible to all users. Data resources placed into this tier should be accompanied by metadata that is also published to ChesapeakeData.
    b. If data are not accessed frequently, the data resource will be moved to the **Infrequent Access Tier** within the Active Data Storage location. This move is done automatically based on access history and is invisible to the user. It is a cost-saving mechanism that is performed by design with no visible change to the data presentation, quality or retrieval time.

2. If your data resources are NOT considered authoritative, published content, please contact the CBPO Data Managers to request access to the CBPO Data Center Strategic Data Plan. This plan dictates how working data should be managed and stored.
3. If your data resource has been considered a **CBPO quality-assured authoritative data resource** but has been superseded OR has not been used within the past seven years, it should be moved into the **Cold Data Archive**. Moving data to the Cold Data Archive involves the following:
    a. Review the Cold Data Archive Migration Checklist
    b. Understand requirements and processes for recovering data that is migrated into the Cold Data Archive.
    c. Contact the CBPO Data Center Managers to request that the content be placed in the Cold Data Archive.

CBPO Data Storage and Archiving

Chesapeake Bay Program
Science. Restoration. Partnership.



Has the data resource ever been considered **authoritative** or used to create published, authoritative derivative products?

NO

OUTSIDE SCOPE
Refer to **CBPO Data Center Strategic Data Plan** for managing non-authoritative data resources

YES

Within the **past seven years,** has the data resource been considered **any** of the following:

- The most current, published, authoritative version of the resource?
- Used by other authoritative resources or derivative products?

NO

Data are moved ONLY after coordination with CBPO Data Center Managers

YES

Annual data resource review

Active Data Storage with Intelligent-Tiering

Frequent Access

Data automatically moved between storage options based on usage

All data in this group are **instantly** accessible

Infrequent Access

Cold Data Archive

Data are accessible within **hours**

Data resources stored in Cold Data Archive:

- Are not returned to Active Data Storage
- Are retained indefinitely
- Require several steps for recovery and access

Figure 1. CBPO Data Archiving Decision Tree

Definitions

- **“Types” Of CBPO Data**
    - **CBPO Quality-Assured Authoritative Data Resources:** Quality controlled and reviewed data resources that are considered the most up to date version of the content. Authoritative resources can be superseded by newer versions of content; however, any data resource that has been classified as “authoritative” at any time during its lifecycle shall follow the processes outlined for authoritative data at CBPO
    - **Non-authoritative Data:** Data resources that are created as interim products that may be used for internal processing or other operations when evaluating, managing, or reviewing data. These interim data resources are not published for use as official, quality controlled authoritative data products for consumers.
    - **Historical Data:** Data that is retained for record-keeping, analysis, or reference purposes. It represents past records or information that are preserved for their long-term value. Historical data may or may not be actively used in day-to-day operations but are considered important for historical context. Historical data can be considered the authoritative, older version of a data resource that represents a specific time period. An example of CBPO historical data resources is the CBPO water quality database, which contains annual versions of water quality data for CBPO watershed from 1984 to present.
    - **Superseded Data:** Data that has been replaced or updated by another data resource, making it no longer the current or authoritative version. This often happens when new data or records are introduced that are considered more accurate, complete, or relevant. An example of superseded CBPO data is an older version of modeling data, which has been superseded by more recent modeling content. Older versions of authoritative modeling data resources should be retained either in active storage or cold storage, based on age and usage.

- **Processes Related to Data Archiving**
    - **Intelligent-Tiering:** a process whereby data are automatically stored in one of several tiers based on access patterns; one tier optimized for frequent access, a lower-cost tier optimized for infrequent access, and a very-low-cost tier optimized for rarely accessed data.
    - **Data Archiving:** The process of moving data to lower-cost storage while ensuring it's accessible when needed. Archived data resources will require more time for retrieval than content stored in active data storage.
    - **Data Versioning:** The creation and management of multiple releases of data, all of which have the same general function, but are improved, upgraded or customized. Different versions of the same data are kept in different places, based on when it was made and how it was changed.
    - **Data Deprecation**: The process of intentionally and systematically discontinuing the use or availability of certain data.
    - **Data Destruction/Disposal:** The process of deleting data that's no longer needed. Data stewards may need to delete data to comply with regulations and/or to save on storage costs.

- **Types of Data Storage**
    - **Active Data Storage:** a storage class for authoritative, published CBPO data that should be available for immediate access. This class of data includes both frequent access storage and infrequent access storage. All data stored within the active data storage tier is immediately accessible to end users.

    - **Frequent Access Storage:** Primary operational storage used for authoritative data resources and derivative products that are considered the most up to date version of the content. This type of storage is designed for data that is accessed or modified frequently, prioritizing high performance and speed. This storage tier may be referred to as "hot" storage and is suitable for data resources that require immediate access.
    - **Infrequent Access Storage:** A storage method for data that is not frequently accessed but is important enough not to be moved to cold storage. This storage class provides price/performance that is cost-optimized for files accessed generally not more than a few times per quarter.   This data, situated between the frequent access tier and the Cold storage archive, is stored in an efficient and cost-effective manner that allows for reasonable access times, generally within seconds.
  - **Cold Data Archive:** A storage method designed for long-term preservation of infrequently accessed data, prioritizing cost-effectiveness and slower access speeds over rapid retrieval. Data stored in Cold storage will typically take several hours to retrieve. Data are moved to this access tier after consultation between CBPO data center managers and data stewards. Data moved to Cold Storage are not returned to Active data storage.

Cold Data Archive Migration Checklist

This simple checklist should be used to ensure that data resources being considered for migration to Cold Data Archive are placed with appropriate documentation, naming conventions, and housekeeping exercises.

- □ The data resource and related products (if any) have not been accessed within the past seven years
- □ The data resource and related products (if any) are not subject to a litigation hold or other records retention requirements.
- □ There are no active web applications, websites, or other public assets that link to the data resource
- □ All content related to the data resource is included in the package that is created for migration (scripts, metadata, readmes, models, or other information that would be required to use the data resource if it is recovered)
- □ The metadata record(s) at ChesapeakeData have been retired (preferred) or updated to reflect that the data resources are no longer available online
- □ The data resource is labelled with an appropriate naming convention that includes version number (where applicable) and the deprecation date.

Summary

Consistent data archiving practices assist CBPO data stewards in ensuring that authoritative data resources are accessible to customers for the appropriate amount of time. It also creates a reliable structure where older, superseded content is moved to proper storage locations using consistent methods. Defining rules and guidance for these processes can help data stewards feel comfortable in moving superseded data out of active storage and into lower cost storage locations for long term placement. These guidelines will help CBPO in managing quality assured, authoritative data sets and implementing consistent lifecycle practices that focus on prioritizing current data resources while retaining superseded data in a way that maximizes accessibility and cost savings.