Chesapeake Bay Program Data Publishing Guidelines

February 14, 2025

Contents

Introduction 3

Purpose 3

Audience 3

Scope 3

Data Publishing Overview 3

Universal Publishing Requirements 5

Contributing to ChesapeakeData 6

Summary 8

Appendix 9

DataHub 9

Overview 9

Examples of Current Processes and Requirements for Publishing Content 9

Future Considerations 10

ChesapeakeProgress 10

Overview 10

Outline of Current Processes and Requirements for Publishing Content 11

Future Considerations 12

Related Publishing Environments 12

Introduction

Data resources are foundational to the success of the Chesapeake Bay Program Office (CBPO), with effective management, quality control, publication, and use critical to its mission. A carefully formulated framework of data publishing requirements and processes helps stewards maximize resource usefulness for stakeholders. The Data Governance Workgroup has prepared this set of guidelines to promote consistency and to provide clarity for those publishing data to CBPO environments. Please refer questions to the CBPO data managers.

Purpose

This document presents the requirements for publishing to CBPO data-sharing environments. It focuses on **Universal Publishing Requirements** for all stewards and environments and on specific instructions for adding content to ChesapeakeData; addresses requirements for publishing to DataHub and ChesapeakeProgress; and explains relationships between the three environments and certain others used to distribute CBPO data resources, such as the CBPO website, ScienceBase, and the Chesapeake Assessment Scenario Tool (CAST). It is a living document that will be updated as data and technology resources at CBPO evolve.

Audience

The primary audience includes CBPO data managers, stewards, and analysts who work with CBPO authoritative data that may be published to CBPO data-sharing environments. The material presented is also intended for those working with CBPO quality-assured data who may post content to the CBPO website and other public-facing CBPO websites.

Scope

The requirements 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 via direct Chesapeake Bay Program (CBP) funding or indirect (matching) funds. Authoritative data may be public or internal, current or legacy, and raw or processed. Non-authoritative data, while often crucial in supporting CBPO's mission, are outside the scope of this document. Non-authoritative data resources include those obtained from external collaborators and data generated internally that have not undergone a documented quality assurance and quality control (QA/QC) process.

Data Publishing Overview

CBPO's data publishing portfolio is made up of the following core systems: a **central metadata repository,** called ChesapeakeData; a **data access tool** for CBPO's environmental monitoring data, called DataHub; and a **data presentation and access tool** for CBPO's environmental indicators data, called ChesapeakeProgress.

- **ChesapeakeData** contains metadata records. Metadata records in the ChesapeakeData catalog link users to data contained in DataHub and ChesapeakeProgress, as well as to resources stored elsewhere, such as those posted to ScienceBase and the CBPO website. ChesapeakeData also provides users with the ability to link to data stored within its own environment.
- **DataHub** provides access to authoritative, reviewed, downloadable files for monitored water quality, living resources, toxics, and fluorescence data. Data are consistently and rigorously quality-assured and documented.

- ChesapeakeProgress provides access to authoritative, reviewed environmental indicators data in Microsoft Excel format (along with visual formats like charts and reports), including abundant life, clean water, conserved lands, engaged communities, and climate change.

The three environments listed above are connected by metadata records placed in ChesapeakeData. When stewards create or update content in DataHub or ChesapeakeProgress, they must create or update a related record in ChesapeakeData so that information is discoverable across the CBPO data landscape. Similarly, when spatial data are published to the CBPO website or contributed to ScienceBase, a corresponding ChesapeakeData metadata record must be created or maintained. (Please see section "Contributing to ChesapeakeData" below for details on the process for publishing metadata to ChesapeakeData.)

Some examples (with links) of data resource types tracked by metadata published to ChesapeakeData:

- **DataHub:** ChesapeakeData contains a metadata record that points to the data made available through DataHub.
- **ChesapeakeProgress:** ChesapeakeData contains a metadata record that describes "Changes in Stream Temperatures" data posted on ChesapeakeProgress.
- **CBPO Website:** ChesapeakeData contains a metadata record that points to the State of Chesapeake Forests 2.0 Chesapeake Bay StoryMap that is made available from the CBPO website
- **ScienceBase:** ChesapeakeData contains a metadata record that points to NHDPlus SPARROW model data made available at ScienceBase.
- **Chesapeake Assessment Scenario Tool (CAST):** ChesapeakeData contains a metadata record that points to the data made available through CAST.

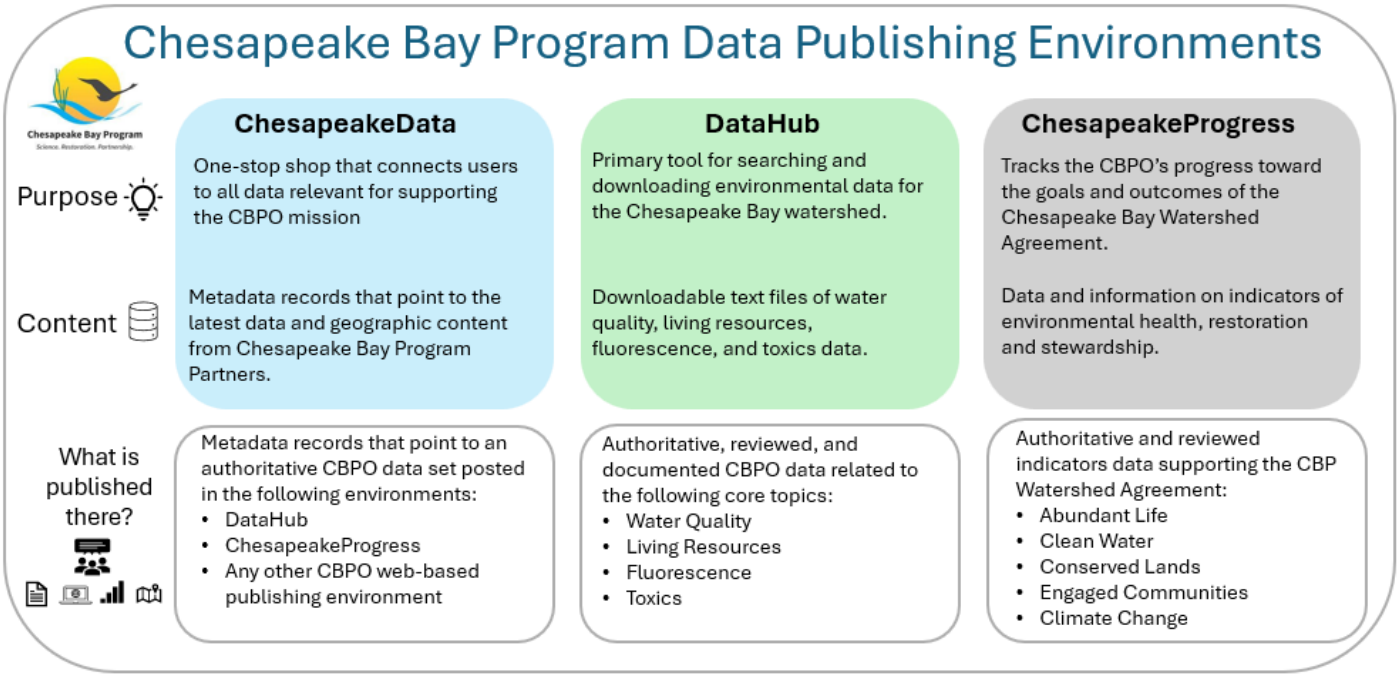

Figure 1. CBPO Data Publishing Environments

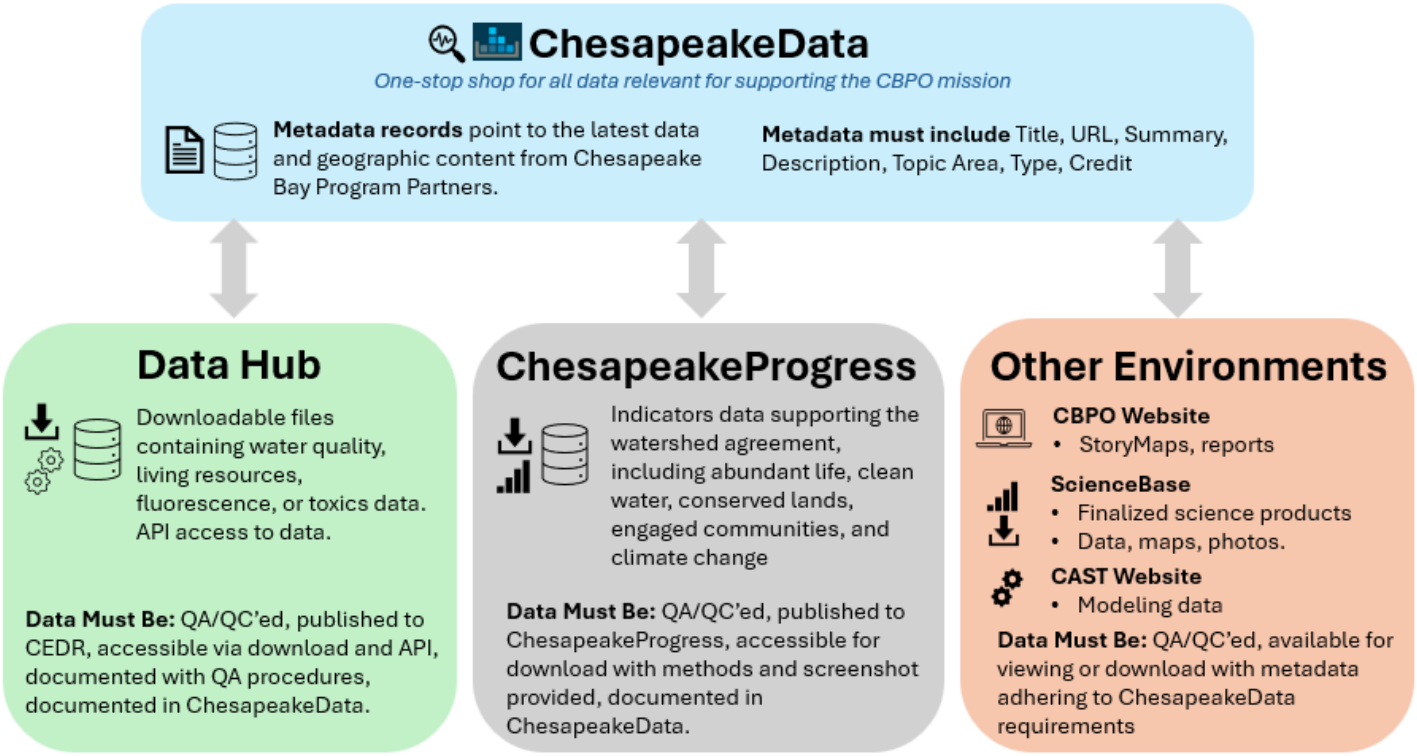

Figure 2. Relationships Between CBPO Data Publishing Environments

Universal Publishing Requirements

When publishing data to CBPO's data-sharing environments, stewards should consider both the proposed **content** and its **quality**, determining whether the content should be published and whether the data meet the publishing requirements, respectively. If the content is appropriate for the selected publishing environment, then the data must meet the quality requirements for that system. Additionally, data should be managed, maintained, and retired according to CBPO's retention and archival procedures (forthcoming). Stewards publishing content to CBPO data-sharing environments should review the following to ensure that data meet the **basic criteria** for doing so.

- Which data **should** be published?
  - Any **dataset funded through grants** should be published to a CBPO data-sharing environment and cataloged in ChesapeakeData.
  - Any **unique dataset** used at CBPO should be published in a CBPO data-sharing environment and cataloged in ChesapeakeData.
  - Any **dataset published to DataHub or ChesapeakeProgress** should have a metadata record published to ChesapeakeData. Data posted to other CBPO environments, such as StoryMaps published to the CBPO website or data published by CBPO personnel to ScienceBase, should also have a metadata record published to ChesapeakeData.
- What are the **minimum requirements** for publishing data?
  - Published datasets must be **relevant to the CBPO mission**.
  - Published datasets must be **unique**, meaning not an exact copy of an existing resource.
  - Published datasets must have **documented quality control processes**.

    - Published datasets must be provided with **documentation** that minimally adheres to ChesapeakeData metadata requirements.

Data meeting these publishing requirements should be made accessible to users online. Options for online dissemination include DataHub, ChesapeakeProgress, the CBPO website, and ScienceBase. When data are successfully posted to an online environment, metadata records **should be contributed to ChesapeakeData** following the steps outlined in the next section. (Please see "Contributing to ChesapeakeData" below.)

Data stewards should review ChesapeakeData catalog content for their published data no less frequently than annually. Anytime published data is changed, related metadata records should be updated accordingly. If a data resource is superseded by a new version or no longer needs to be published, it should be retired. **Guidelines for data versioning and retirement are being developed.** Until those are available, the following questions can help guide stewards in determining if data are good candidates for retirement/archive.

- When should data be **retired**?
    - Published data that are no longer being used should be identified as candidates for archive**.** Specifics of archiving/versioning processes will be decided on a case-by-case basis by the data stewards of the resources. The following should be considered by stewards when determining which content to archive, future availability, and archive processes:
        - Should CBPO continue to pay for the data or archive them?
        - Does CBPO have an obligation to continue to serve the data resources or make them available upon request?
        - Is the data resource still being maintained?
        - Has the data resource been replaced by another dataset?
        - Are the data resources or the metadata incomplete or factually inaccurate?
        - Are there usage and/or download statistics indicating that the dataset is not being extensively utilized?
    - When datasets are archived, all corresponding metadata records in ChesapeakeData should be archived.

Contributing to ChesapeakeData

The ChesapeakeData metadata catalog serves as a central point of access for data, decision-support tools, and other data resources important to the work of the CBPO. Metadata records connect users to data via URLs that point to downloads, maps, reports, APIs, and other delivery endpoints.

Guiding philosophy and principles:

- Providing access to data resources is the overall goal; providing access to custom queries or subsets is not.
- The catalog is the most useful for all audiences when content duplication is minimized.
- National datasets should have a Chesapeake-specific purpose, e.g., a connection to the Chesapeake Bay Watershed Agreement.
- All ChesapeakeData content should be reviewed **annually** and maintained to ensure high-quality content with valid links to the most recent data.

The checklist below presents the basic requirements for content published to ChesapeakeData. It should be revisited annually for each published metadata record. When data of interest meet the given requirements, use ChesapeakeData's Content Suggestion Form to propose inclusion of the data in the catalog. ChesapeakeData stewards will review each submission, verify completeness and fitness for publishing, and provide guidance for

finalizing and releasing content to the public. Figure 3 shows the process generally followed. Please direct questions to data@chesapeakebay.net.

Checklist of Mandatory ChesapeakeData Publishing Requirements			
Criteria	**Yes**	**No**	**Details**
Are your data relevant to CBPO's mission and scientific needs?			
Have you reviewed ChesapeakeData to make sure the exact same content is not already published?			
Is the dataset unique? • Not a copy of another dataset in ChesapeakeData • Not a subset of another dataset, unless significant alterations make the dataset meaningful to stakeholders. • Not simply "clipped" (from a larger dataset) to the Chesapeake Bay area with no other changes			
Does your dataset have metadata?			If the dataset has no metadata, can metadata be created with adequate details?
Do your data link to a resource that is accessible to public users?			
Have your data been reviewed using a documented quality control process that is accessible to users?			
Have your data been reviewed for completeness?			
Is accurate contact information provided with your data?			
Is your dataset the most recent version of the data?			Only the most recent versions of data should be included unless previous versions are needed for current management objectives or have value for subject matter experts.

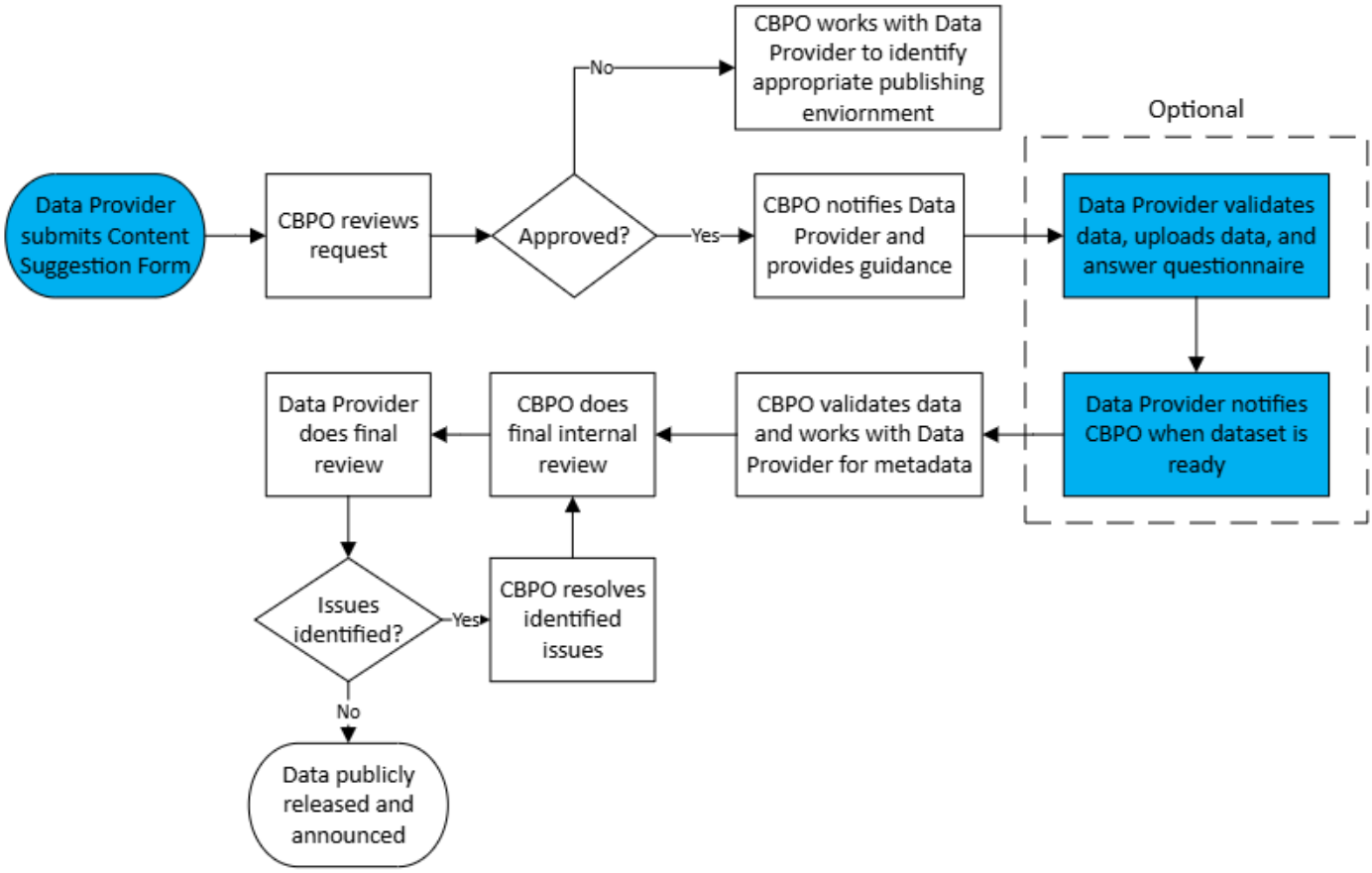

Figure 3. CBPO Data Publishing Process Flowchart

Summary

These guidelines are part of a larger CBPO Data Governance effort. The processes described herein reflect the core principles of the CBPO Data Governance Group, including:

- Discovery and Accessibility
- Classification
- Quality
- Data Standards
- Value
- Transparency
- Interoperability

It is a living document that will be updated as data and technology resources at CBPO evolve. Questions can be directed to: datamanagers@chesapeakebay.net.

Appendix

DataHub

Overview

CBPO's DataHub is a tool for accessing environmental monitoring data for the Chesapeake Bay watershed. It provides users with access to a data download interface and an Application Programming Interface (API) for extracting data. Current data resource types are Water Quality, Tidal Plankton, Tidal Benthic, Nontidal Benthic, Fluorescence, Nutrient Point Source, and Toxics. Data provided are derived from multiple state partners. The content is rigorously reviewed and verified by DataHub stewards. All data are provided with substantial documentation.

Contribution requirements and discipline:

- Data are considered authoritative, have been reviewed, have been verified using a repeatable process, and have undergone rigorous checks to ensure accuracy.
- Data must meet all minimum requirements established in DUET and CEDR.
- Data must include metadata that meets minimum requirements for the resource type (e.g., Water Quality) being published.
- Data must have metadata records posted to ChesapeakeData.

Examples of Current Processes and Requirements for Publishing Content

For more detail, please visit the Water Quality Specifications and the Living Resources Specifications.

Process for Ingesting Data:

- Water Quality and Living Resources data are contributed to DataHub through the Data Upload and Evaluation Tool (DUET). Some data are contributed directly by state data stewards; other data are emailed to CBPO data stewards and then uploaded to DUET.
- Once data pass validation requirements in DUET, they are ingested into the Chesapeake Environmental Data Repository (CEDR).
    - CEDR is an enterprise database that houses all CBPO Water Quality and Living Resources (Tidal Benthic and Nontidal Benthic macroinvertebrates, Tidal Plankton) data holdings since 1984.
- Data are accessible through DataHub once they are published in CEDR. DataHub provides a simple web interface and API for accessing the data.
- The DUET User Guide provides an overview of requirements for contributing to DUET.

Detailed Publishing Steps: Living Resources

- There are three Living Resources data streams: Tidal Benthic, Nontidal Benthic, and Tidal Plankton.
- Information is either emailed to or downloaded by DataHub stewards.
- Data are subject to an initial QA/QC requirement before sending to a DataHub steward.
- Data are transposed into the required format for uploading into DUET.
- DUET performs a series of QA/QC checks in addition to those done by the data provider before data are submitted to the DataHub steward.
- Data that pass all quality checks are pushed to the CEDR database.

- Content is generally emailed annually or twice a year. Varies by state and subject matter area. Nontidal Benthic has a data call every five years.
- Data are delivered under the terms of grants, and some is provided voluntarily.
- Currently, the is no consistent process in place for contributing to ChesapeakeData.

Detailed Publishing Steps: Water Quality

- All submissions directly into DUET.
- Data are submitted as Microsoft Access databases with seven tables.
- Data are run through a set of 160-170 checks.
- After data pass checks, the DataHub steward works with the provider to work through nonfatal errors.
- Upon reaching agreement regarding errors, data are contributed to CEDR.
- Data are categorized as Tier I, II, and III.
- The frequency of data submissions depends on the state and the program.
- Currently, there is no consistent process in place for contributing to ChesapeakeData .

Follow the link to publishing requirements for DUET for more information.

Future Considerations

As of the writing of this document, planning for a modernized DataHub environment is underway. Requirements and priorities are being identified. Questions to be considered include the following:

- *Why do some things go into DataHub, and why don't some others?*
- *What criteria should be used for publishing new data to the redesigned DataHub?*

Example of 2003 Shallow Water Monitoring Data:

Maryland and Virginia provide this data. Each state provides access via its own website. Some analysts have asked if CBPO could house the data for convenient access to the entire dataset. Questions include:

- *How would CBPO house the data?*
- *What would CBPO do with it?*
- *Where would it be stored?*

Answers to questions like these will help shape an implementation supporting a broader collection of monitoring data.

ChesapeakeProgress

Overview

CBPO's ChesapeakeProgress provides access to environmental indicator data resources relevant to meeting the goals of the Watershed Agreement. The web-based interface allows users to view data, download data files, and access documentation regarding methods used to develop the data. CBPO managers rely on environmental indicators to summarize pollution sources, ecosystem conditions, and trends toward meeting specific restoration goals.

The primary pages on the ChesapeakeProgress site have a one-to-one relationship with the goals and outcomes of the Watershed Agreement. There are two kinds of data.

1. Indicator data that track progress toward outcomes.
2. Non-indicator datasets that illustrate factors influencing outcomes. (For example, watershed population is important to those working to achieve the "Land Use Methods and Metrics Development" outcome of the "Land Conservation" goal.)

Outline of Current Processes and Requirements for Publishing Content

For more detail, please visit the ChesapeakeProgress website.

Data Sources:

- Datasets are compiled by partner organizations. Some are multi-partner datasets, and some are not.
- Because a revision to the Watershed Agreement is pending, the list of indicators data is frozen for 2025. New indicators will be adopted/added in response to updates to the Watershed Agreement.
- In many cases, multiple state and federal partners provide data, sometimes funded by grant dollars.
- Workgroup and Goal Implementation Team (GIT) chairs often provide some level of oversight across multiple states collecting the data.
- ChesapeakeProgress utilizes much of the water quality data that appears in other CBPO data products.
- Types of data include quantitative data (numbers, measurements, charts, values shown over time, etc.), geospatial datasets, and social science indicator data, such as results of surveys conducted by GIT workgroups. Certain social science indicators are based on discrete datasets that are relatively recent in origin. ("Local Leadership" is an outcome with some relatively new supporting datasets.)

Quality Control Processes:

- Analysis and Methods (A&M) documents are provided with every dataset. (See example here.) They are completed in a collaborative way.
- A&M information may not change substantially from one year to the next.

Data Deliverables:

- Underlying data are downloadable. The spreadsheet option may be the simplest representation. Sometimes there is a more complex underlying dataset. In some cases, raw data resides in the states, and the CBPO has a simplified version.
- Some data have associated grant funding, and some do not. Related questions and considerations:
    - When not funded, are there implications for the delivery of this data in the future?
    - Does lack of funding impact quality or availability?
    - What incentive is there for states to provide data in the absence of grant funding? (As an illustration, state identification of healthy watersheds was a one-time effort. It was the best that could be done at the time. That data was prioritized for a short period, then no longer provided and/or updated.)

Update Frequency:

- Update frequency varies. Annual in most cases. Some datasets have established baselines and/with no updates since. Some data obtained from human surveys may have periods as long as five years. Most are annual.

Future Considerations

The Watershed Agreement is being revised in 2025. The updates will impact which indicators are measured and evaluated, as well as the data needed. For long-term consistency of indicator data, publishing leads should work to influence indicator data delivery under the new Watershed Agreement. When the CBPO has established new goals and identified methods for measuring them, language should be included to ensure that raw data and analytical formulas accompany processed data.

- Grant guidance should be written to make sure that **datasets** needed for ChesapeakeProgress indicators are provided over the long term. Could the Data Governance Workgroup provide language so that future grants require that the CBPO receives **raw data** and is included in the **entire flow** of the data lifecycle for information being measured?
- In the past, some outcomes have been chosen and defined without clear indicators in mind. Would be worthwhile for the Data Governance Workgroup to help ensure that goals and outcomes are measurable, with real underlying data.

In addition to providing language to ensure delivery of the full set of data and methods that support indicator progress measurement and tracking, CBPO should revisit and reinstate the process for **proposing new indicators:**

- What are the decision rules for publishing new data to ChesapeakeProgress? When new indicators are proposed, the lead should ensure that requirements are written so that users can download and access all data, methods, and documentation needed for understanding how indicator status was derived.
- Traditionally, the Status and Trends Workgroup has overseen proposals for new indicators, but none are being considered. Further, the CBPO Indicators Coordinator served as CBPO's gatekeeper and manager for all indicators, but that role is currently vacant. The proposal process and the Indicators Coordinator role should be reestablished when the new Watershed Agreement is in place. The Indicators Coordinator should follow the processes and rules established by the Data Governance Workgroup.

Related Publishing Environments

The websites listed below provide access to CBPO resources and data. In many cases, content on these sites should be published to ChesapeakeData, but not in all cases. General guidelines given here help determine which content should be included on ChesapeakeData. Data stewards who post content to these sites should work with ChesapeakeData stewards to determine which assets to include in the metadata catalog per the "Contributing to ChesapeakeData" section above.

- **Data Published to CBPO Websites:** StoryMaps, reports, charts, and other resources that provide access to data related to conditions in the Chesapeake Bay Watershed should have metadata records published to ChesapeakeData.
- **Data Published to the ScienceBase Website:** All officially published datasets that are developed by USGS scientists as part of their work supporting the CBPO should have metadata records published to ChesapeakeData.
- **Data Published to the Chesapeake Assessment Scenario Tool (CAST) Website:** Modeling data published to CAST should have metadata records published to ChesapeakeData.