**SECTION 13.05 – Quality Assurance/Quality Control Approaches for Existing Scientific Data and Technical Information**

**TABLE OF CONTENTS**

[POLICY STATEMENT 2](#_Toc512933671)

[REFERENCES 2](#_Toc512933672)

[BACKGROUND 3](#_Toc512933673)

[Figure 1. A Simple Graphical Summary of the Literature Screening Process 4](#_Toc512933674)

[DEFINITIONS 4](#_Toc512933675)

[Existing Data and Information 4](#_Toc512933676)

[Grey Literature 5](#_Toc512933677)

[Metadata 5](#_Toc512933678)

[Informative or Pertinent Studies 5](#_Toc512933679)

[The Agency’s Five General Assessment Factors 5](#_Toc512933680)

[RESPONSIBILITIES 9](#_Toc512933681)

[Principal Investigators (PI) 9](#_Toc512933682)

[QA Managers and Directors of QA 9](#_Toc512933683)

[Supervisors 9](#_Toc512933684)

[PROCEDURES 9](#_Toc512933685)

[DELEGATIONS 23](#_Toc512933686)

[APPROVAL 23](#_Toc512933687)

[Appendix A. Existing Data for Addressing Research Questions & Data Limitations 24](#_Toc512933688)

[Appendix B: Collecting, Screening, Evaluating, & Selecting Existing Data 26](#_Toc512933689)

[Figure 2: General Process 26](#_Toc512933690)

[B1. General Considerations for Existing Data (Validation, Analysis, Review) 27](#_Toc512933691)

[B2. Systematic Review in EPA’s IRIS Assessment Development Process 29](#_Toc512933692)

[B3. Evaluating the Quality of Individual Studies for EPA’s Integrated Scientific Assessments 30](#_Toc512933693)

[Appendix C. General Evaluation Criteria for Existing Data vs. Measurement Data 31](#_Toc512933694)

[Appendix D. SAMPLES & EXAMPLES FOR EXISTING DATA 33](#_Toc512933695)

[D1. Boilerplate QA Language for Extramural Performance Work Statements 33](#_Toc512933696)

[D2. Sample QAPP Template 34](#_Toc512933697)

[D3. Example QAPPs Available Online 34](#_Toc512933698)

[D4. Guidance and Checklists 35](#_Toc512933699)

[D5. Plans, Literature Search Strategy, and Preliminary Evidence Tables 35](#_Toc512933700)

[D6. Using Genomics data 35](#_Toc512933701)

[D7. Considerations for GIS Datasets (Stream Cat) 36](#_Toc512933702)

[ADDITIONAL REFERENCES 37](#_Toc512933703)

# POLICY STATEMENT

ORD scientists routinely rely upon existing scientific data and technical information. This section of the ORD Policy and Procedure Manual (PPM) outlines quality assurance (QA) and quality control (QC) procedures to ensure transparency and objectivity in the collection and use of existing data and provides an approach for the development of a Quality Assurance Project Plan (QAPP) and other project and product QA documentation to ensure that data quality meet project objectives. This policy does not create a new requirement for QAPPs for research involving the use of existing data (see EPA QA/R-5 and EPA QA/G-5). The impetus in providing this PPM stems from the Agency’s increasing reliance on making use of existing information, the corpus of which is vast. An approach for systematically reviewing existing information is presented but not all presented quality considerations will apply to all research projects. The considerations outlined are provided for informational purposes only; i.e., rather than serve as a prescriptive template. This guidance is to be considered when developing or reviewing QA planning documents.

# REFERENCES

1. EPA QA/G-5, “Guidance for Quality Assurance Project Plans, Chapter 3, pp. 47-57. Projects Using Existing Data,” Environmental Protection Agency, Office of Office of Environmental Information, Washington, DC, (EPA/240/R-02/009), December 2002. *<https://www.epa.gov/quality/guidance-quality-assurance-project-plans-epa-qag-5>*
2. EPA QA/R-5, “EPA Requirements for Quality Assurance Project Plans,” Environmental Protection Agency, Office of Environmental Information, Washington, DC, (EPA/240/B-01/003), March 2001.[*https://www.epa.gov/sites/production/files/2016-06/documents/r5-final\_0.pdf*](https://www.epa.gov/sites/production/files/2016-06/documents/r5-final_0.pdf)

(3) 67 FR 8452, “Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility and Integrity of Information Disseminated by Federal Agencies,” Office of Management and Budget, Federal Register, February 22, 2002.   
[*http://www.gpo.gov/fdsys/pkg/FR-2002-02-22/pdf/R2-59.pdf*](http://www.gpo.gov/fdsys/pkg/FR-2002-02-22/pdf/R2-59.pdf)

(4) EPA 100/B-03/001 “A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information (2003) and Addendum (2012),” Science Policy Council, Washington, DC. [*https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information*](https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information)

(5) EPA/100/B-15/001, U.S. Environmental Protection Agency Peer Review Handbook, 4th Edition, Science Policy Council, Washington, DC, 2015.   
*https://www.epa.gov/sites/production/files/2015-10/documents/epa\_peer\_review\_handbook\_4th\_edition\_october\_2015.pdf*

(6) EPA PPM 13.07. Office of Research and Development, Policy and Procedure Manual, Chapter 13.7, Use of the Graded Approach for Quality Assurance of Research, 2017.

# BACKGROUND

In keeping with Office of Management and Budget guidelines (67 FR 8542), EPA’s quality system requires that all of the Agency’s environmental data be of known and documented quality. It is relatively straightforward to characterize the quality of a project’s newly generated data. It is more challenging to characterize the quality of existing data and information previously collected for other projects or intended applications. The rigor applied in examining existing data is proportional to the importance of the study (Use of the Graded Approach for Quality Assurance of Research; see PPM 13.07 [2017]). In other words, the level of review and stringency of evaluation criteria will depend upon projected use of the existing data.

There are numerous benefits for using existing data and information to meet project objectives. Relative to the collection of new data, using existing data and information may be advantageous to attaining project objectives for a variety of reasons, including:

* cost effectiveness,
* ready availability and/or ease of collection,
* better technology now exists for analysis of historical data
* constraints on producing new data (e.g., access to site is restricted), and
* provision for more detailed and/or exhaustive information (e.g., covering a broader geographic distribution or longer time span).

By taking advantage of the strengths of existing data and information while acknowledging potential limitations on their use, researchers can maximize project resources in a transparent manner. The purpose of this PPM is to provide guidance about QA/QC considerations for the collection and use of existing scientific data and information for all ORD research activities. Although this document primarily focuses on existing data and information from published scientific literature, it also addresses other potential sources of existing data such as, geospatial data, and public-use and non-public-use datasets.

Figure 1 illustrates a graphical summary of the literature screening process. For simplicity’s sake this is presented as a linear process but in reality researchers may iterate through searching, screening and considering papers many times before information is included as project data.

****

## Figure 1. A Simple Graphical Summary of the Literature Screening Process

(Source: *Integrated Science Assessment Preamble,* [*https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310244*](https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310244))

The Integrated Science Assessment (ISA) Preamble describes the process of searching the literature, selecting studies for consideration, evaluating study quality, synthesizing and integrating the evidence, and characterizing the evidence for public health and welfare impacts of criteria air pollutants, (see Appendix B3).

# DEFINITIONS

## Existing Data and Information

Scientific data and technical information that a researcher plans to use but that have not been newly generated by the current project are called “existing data and information.” They may also be known as secondary data, literature data, non-direct measurements, or 3rd party data and information. Sources for existing data and information may include journal articles, technical reports, model output, large electronic data sets and databases.

## Grey Literature

Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by libraries and institutional repositories, but not controlled by commercial publishers; i.e. where publishing is not the primary activity of the producing body (12th International Conference on Grey Literature; Prague, Czech Republic 2010).

## Metadata

Metadata (data about data) is the information about the dataset or study itself to facilitate its use and understanding. Metadata describes why and how the data were collected and its limitations. Specifically, metadata for existing data and information should include source, study quality, data quality, author, methods, analytics (e.g., statistical methods), results, summaries (e.g., abstract), and other criteria as appropriate. Metadata may also include data collection, tabulated data summaries, standards, documented QA/QC procedures, and other project-specific documents. Metadata can help determine the value of existing data; namely, whether data are accurate, reliable, available, and fit for use in a project for which data were not originally intended.

## Informative or Pertinent Studies

Research that represents new advances confirming previous informative or pertinent studies, replaces weaker previous studies, or is otherwise the best or only data available.

## The Agency’s Five General Assessment Factors

When evaluating the quality and relevance of scientific and technical information, the considerations that the Agency typically takes into account can be characterized by five general assessment factors (<http://www.epa.gov/spc/assess.htm>). Example questions that could be raised by the consideration of each of the five factors for various types of information are provided for each assessment factor. Given the very general nature of these assessment factors, the Agency felt that a compilation of such illustrative questions would most clearly convey the intended nature and breadth of the assessment factors, and how they would be reflected in an evaluation of various types of information. However, the applicability of these factors depends on the individual situation, and researchers should exercise discretion to consider and use factors and approaches on a case-by-case basis that may differ from the illustrative considerations presented below:

1. **Soundness**

The extent to which the scientific and technical procedures, measures, methods or models employed to generate the information are reasonable for, and consistent with, the intended application.

Example questions:

1. Is the purpose of the existing study or data set reasonable and consistent with the design of the intended application?
2. To what extent are the procedures, measures, methods, or models employed to develop the information reasonable and consistent with sound scientific theory or accepted approaches?
3. How do the study or data set’s design and results compare with existing scientific, statistical, or economic theory and practice? Are the assumptions, governing equations and mathematical descriptions employed scientifically, statistically, and technically justified? Is the study or data set based on sound scientific, statistical, or econometric principles?
4. In the case of a survey, have the questionnaires and other survey instruments been validated (e.g., compared with direct measurement data)? Were checks for potential errors made during the interview process?
5. How internally consistent are the existing study or data set’s conclusions with the data and results presented?
6. Evaluate data quality relative to the intended use of the data. Identify criteria or metadata for judging data or information quality, including accuracy, precision, representativeness, completeness and comparability. Do individual data points fall within known or biologically plausible ranges?

2. **Applicability and Utility**

The extent to which the information is relevant for the Agency’s intended use.

Example questions:

1. How useful or applicable is the scientific, statistical, or economic theory applied in the study or data set to the Agency’s intended use of the analysis?
2. How relevant are the study or data set’s purpose, design, outcome measures and results to the Agency’s intended use of the analysis (e.g., for a chemical hazard characterization)?
3. Relative to the Agency’s intended use of the data, how unique or distinct is the data set? Would it be possible to generate a similar, equally useful data set under the Agency’s internal information quality systems, and if so, what would be the cost-benefit for using the external data set versus the Agency’s internally generated data set?
4. Are the domains (e.g., duration, species, exposure) where the model or results are valid useful to the Agency’s application?
5. How relevant is the study or data set to current conditions of interest? For example, in the case of a survey, are conditions likely to have changed since the survey was completed (i.e., is the information still relevant)? Is the sampled population relevant to the Agency’s current application? How well does the sample take into account sensitive subpopulations?
6. What constraints are there on the usage of the information or data (e.g. legal, logistical, programmatic, privacy/confidentiality, language barriers, etc.)?

3. **Clarity and Completeness**

The degree of clarity and completeness with which the data, assumptions, methods, quality assurance, sponsoring organizations and analyses employed to generate the information are documented.

Example questions:

1. To what extent does the documentation clearly and completely describe the underlying scientific, statistical, or economic theory and the statistical and analytic methods used?
2. To what extent have key assumptions, parameter values, measures, domains and limitations been described and characterized?
3. To what extent are the results clearly and completely documented as a basis for comparing them to results from other similar tests?
4. If novel or alternative theories or approaches are used, how clearly are they explained and the differences with accepted theories or approaches highlighted?
5. Is the complete data set accessible, including metadata, data-dictionaries and embedded definitions (e.g., codes for missing values, data quality flags and questionnaire responses)? Are there confidentiality issues that may limit accessibility to the complete data set?
6. In the case of a modeling exercise, have the definitions and units of model parameters been provided? To what extent have the procedures for applying the model been clearly and completely documented? How available and adequate is the information necessary to run the model computer code?
7. To what extent are the descriptions of the study, data set, or survey design clear, complete and sufficient to enable the study or survey to be reproduced?
8. Have the sponsoring organization(s) for the study/information product and the author(s) affiliation(s) been documented?

i) To what extent are the procedures for quality assurance and quality control of the data documented and accessible?

4. **Uncertainty and Variability**

The extent to which the variability and uncertainty (quantitative and qualitative) in the information or in the procedures, measures, methods or models are evaluated and characterized.

Example questions:

1. To what extent have appropriate statistical techniques been employed to evaluate variability and uncertainty? To what extent have the sensitive parameters of models been identified and characterized?
2. To what extent do the uncertainty and variability impact the conclusions that can be inferred from the data and the utility of the study? What are the potential sources and effects of error and bias in using these data for the subject research?

c) Did review of the study or data set identify potential uncertainties such as those due to inherent variability in environmental and exposure-related parameters or possible measurement errors?

5. **Evaluation and Review**

The extent of independent verification, validation and peer review of the information or of the procedures, measures, methods or models.

Example questions:

1. To what extent has there been independent verification or validation of the data set or study method and results? What were the conclusions of these independent efforts, and are they consistent?
2. To what extent has independent peer review been conducted of the data set or study method and results, and how were the conclusions of this review taken into account?
3. Has the procedure, method or model been used in similar, peer reviewed studies? Are the results consistent with other relevant studies?

d) In the case of model-based information, to what extent has independent evaluation and testing of the model code been performed and documented?

# RESPONSIBILITIES

## Principal Investigators (PI)

PIs are responsible for working with quality staff to develop and follow a QAPP that will either separately or as part of a project QAPP (a) outline a procedure for documenting the approach, basis for inclusion/exclusion, and resources to be evaluated for existing data collection; (b) document the criteria for critically evaluating adequacy of existing data for meeting project objectives and documenting the evaluation results; (c) document project metadata; and (d) document any limitations or potential biases from the use of existing data on project conclusions.

## QA Managers and Directors of QA

QA Managers are responsible for assessing the implementation of this policy for projects based on existing data and information and for reviewing and approving QA documents, such as QAPPs and any other related project QA documentation (e.g., QA section of a draft or final product as a result of the project.) Directors of QA are responsible for assessing the implementation of this policy at the program level.

## Supervisors

Supervisors are responsible for ensuring that PIs are familiar with this policy and for providing oversight in implementing this policy.

# PROCEDURES

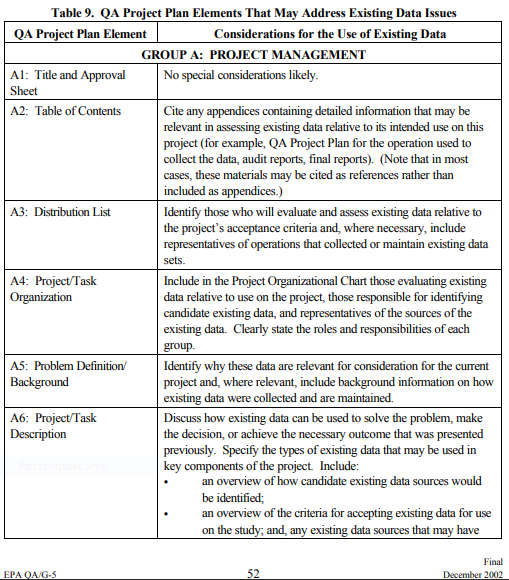
This PPM outlines considerations for developing and reviewing QAPPs and other related QAQC documentation for projects using existing scientific and technical data and information. This PPM applies to existing data and information for a project that may be from a single source or agglomerated or multiple sources and intended for use in a meta-analysis. The outline below is cross-referenced to the Agency’s guidance for QAPPs, EPA QA/G-5. This section follows the standard EPA QA/G-5 QAPP structure as indicated.

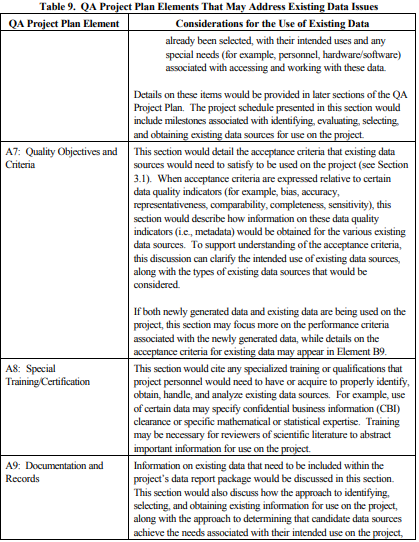
**Using EPA QA/R-5 QAPP Elements**

The EPA QA/R-5 QAPP elements in EPA QA/G-5, Chapter 3, Table 9, pp. 52-57, have been copied into this document, below, to help you prepare and review QAPPs for existing data that adhere to EPA QA/R-5. **Generally, all EPA QA/R-5 QAPP elements apply, except B1 through B8 (the Data Generation elements).** However, if your project includes modeling of existing data, you will likely use a combination of QAPP elements from Section B, and include any or all of B1 through B8. They key elements are B9, B10, C1, and D1-3.

The following pages provide the EPA QA/R-5 requirements and EPA QA/G-5 framework and guidance for developing and documenting the use of existing data and information in a QAPP. In this PPM, the QAPP Sections have been copied directly from Table 9 in the Agency’s guidance and displayed as three parts: QAPP Section A, Section B, and Sections C & D (combined). You will find additional notes below each section’s table that provides further detail.

**QAPP SECTION A**





**Additional notes on Section A (above) for project management, objectives, organization, and responsibilities. The QAPP should include:**

A4: Responsibilities of ALL or MOST project participants are documented in this section

* Key personnel and their organizations
* Designation of responsibilities for planning, coordination, data gathering, data analysis, report preparation, and quality assurance

A5: The research question is clearly stated and provides detailed background information

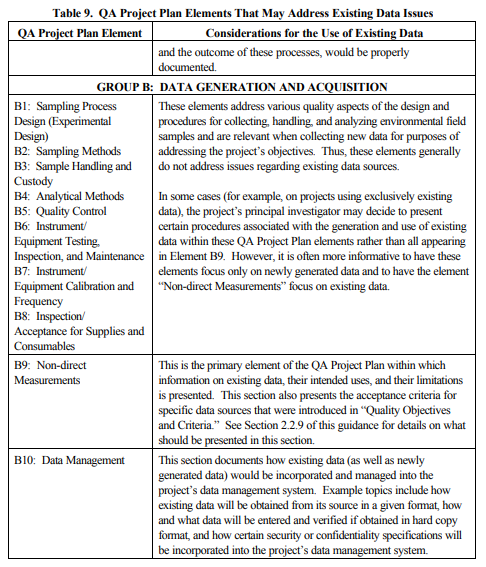
A6: Project planning deliverables are mentioned, such as:

* Schedule
* Interim report(s)
* Expected final product (e.g., journal article, final report, etc.)

A7: A clear statement about how existing data will be used to address research questions in the form of attainable project objectives

* Specifies the desired quality of existing data;
* Specifies the needed format of the existing data
* Identifies necessary data conversion and validation that conversion is accurate

**QAPP SECTION B**



**Additional notes on Section B9-10 (above) for Non-direct Measurements, (**data acquisition, collection, evaluation, synthesis, and data management). The QAPP should include, as appropriate:

**B9**: **Sources, search strategy, types of existing data, acceptance criteria and limitations are identified in this section.**

* Design approach to capture the maximum amount of relevant information (see Appendix B in this PPM)
* Search strategy, basic or detailed criteria, rationale for acceptance/rejection, and inclusion/exclusion to project and citation tracking.
* Types of existing data and information may include:
  + Environmental numbers, measurements, computed values etc., including background/historical data, obtained from governmental (local, state, tribal, federal, international), academic, and corporate databases and records (e.g., [www.data.gov](http://www.data.gov))
  + Scientific literature published in peer-reviewed journals
  + Database information on genomes, proteomes, transcriptomes, metabolomes
  + “Grey” literature, which includes scientific reports, proceedings from meetings, and the like that may be distributed without a formal review from experienced and objective reviewers
  + Existing sampling and analytical data from reports on previous investigation(s) of an area under consideration
  + Computer model simulations and outputs from existing models
  + Physical conditions of the environment (e.g., meteorological, hydrologic, geologic)
  + Spatial information (survey data, geographic information system data or “layers,” topographical maps, etc.)
  + Photographs
* **Outline process to verify availability of existing data** **for addressing project** objectives
* **Record metadata**, such as, data source, originating organization, report title, type of information, and date
* Identify how and where metadata will be recorded
* **Existing data from sources** **which have well-documented QA/QC procedures will have extensive metadata** with the data and no additional data evaluation may be necessary; for example:
  + Air quality data from EPA's Air Quality System (AQS)
  + Meteorological data from the National Climatic Data Center (NCDC)
  + Demographic and housing data from the Bureau of the Census
  + Mortality data from the National Center for Health Statistics (NCHS)
  + Hospitalization data from the Center for Medicare and Medicaid Statistics (CMS)
  + Clinical data from the National Health and Nutrition Examination Survey (NHANES)
  + Satellite imagery from the National Environmental Satellite, Data, and Information Service (NESDIS)
  + Land use data, which is utilized from the Environmental Systems Research Institute (ESRI) as part of EPA's official GIS platform
  + Standards data from the National Institute of Standards (NIST)
* **Document Plan for Data Review and Analysis**
  + Document project objectives
  + Data/information quality considerations regarding fitness for use. List qualifications and known limitations of existing data
  + Each project will have its own acceptance criteria
* **Determine Whether Data Have Constraints not Covered Under this PPM** 
  + Legal
  + Programmatic
  + Logistical (e.g., limitations on access to existing data)
  + Privacy/confidentiality
  + Language medium (e.g., broad literature surveys may preclude foreign language studies for practical reasons such as limited time/resources available per reference)
* **Establish Criteria for Evaluating Data Quality Relative to the Type of Data and Intended Use** (see Appendix A in this PPM). These items may be documented in QAPP Section B9 or in Sections C & D (repeated in Section C & D in this PPM since some document criteria in either B9 or C & D). Sometimes it is easier to list it all in a single section and reference it later. Section B9 generally documents the criteria for evaluating the overall quality of a study or dataset while Sections C & D document performing the assessments, reviews, and verifying individual data values. To evaluate epidemiological studies and similar, you will want to use the EPA’s Five General Assessment Factors (GAF) described earlier in this PPM. To evaluate measurement-derived datasets such as geospatial data or data generated by a lab or in the field, you will want to inspect data values and apply the criteria below, although you can apply the general Assessment Factors.

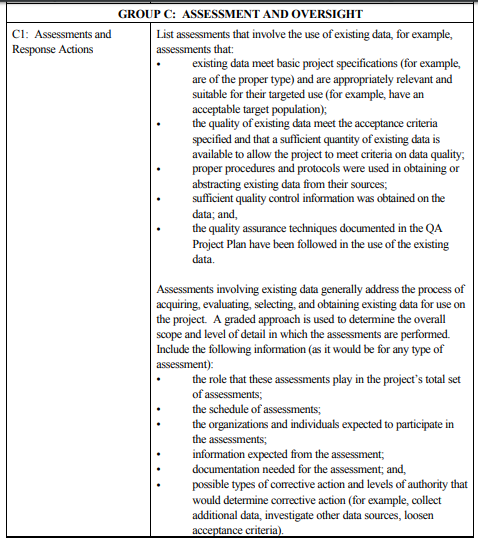
**Example criteria, general & specific**:

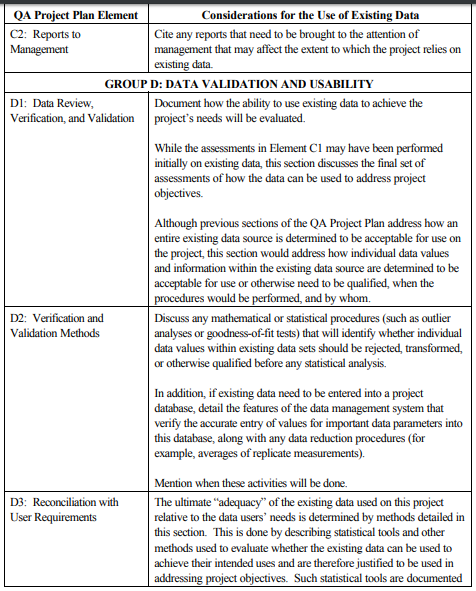
* + Describe procedures for determining the quality of the existing data
  + Identify criteria or metadata for judging data or information quality, including:
  + EPA’s Five General Assessment Factors
    - Soundness
    - Applicability & Utility
    - Clarity & Completeness
    - Uncertainty & Variability
    - Evaluation & Review
  + Measurement Data
    - Accuracy
    - Precision
    - Representativeness
    - Geographic
    - Temporal
    - Technological
    - Completeness
    - Comparability
    - Age of data
    - Samples collected using the same or comparable statistical design
    - Samples or information collected over time at the same location
    - Samples represent similar matrices
    - Data generated using the same or comparable methods or SOPs
    - Data reported in comparable units of measurement
    - Data collected and managed according to an approved EPA R5-compliant QAPP or appropriate project QAPP according to the Graded Approach (PPM 13.7)

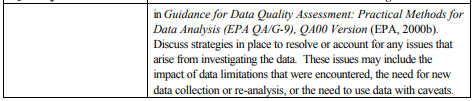
**B10**

* Data management is critical especially if acquiring a significant amount of data. This section should contain enough detail to describe how data will be stored, shared, and secured.

**QAPP SECTIONS C & D**







**Additional notes on Sections C & D (above), on review, verification, and validation.** The QAPP should include, as appropriate:

**C1:**

* **Identify Aspects of Research Resulting in Informative or Pertinent Studies**

When evaluating data from the literature, it is recommended to use general data quality categories which can be customized to meet your project’s objectives. (See above page 5 of this PPM for General Assessment Factors Guidance for the five general factors).

* Develop a data quality checklist to conduct the assessment of the data
* Qualitatively rank results using a data quality checklist
* Quantitatively rank results using a data quality checklist
* Develop a template to document the results of the review

**D1-3:**

* **Criteria for Evaluating Data Quality Relative to the Type of Data and Intended Use** (see Appendix A in this PPM). This information also appears in the Notes for Section B9 in this PPM. These items may be documented in QAPP Section B9 or in Sections C & D. Section B9 generally documents the criteria for evaluating the overall quality of a study or dataset while Sections C & D document performing the assessments, reviews, and verifying individual data values. To evaluate epidemiological studies and similar, you will want to use the EPA’s Five General Assessment Factors (GAF) described earlier in this PPM. To evaluate measurement-derived datasets such as geospatial data or data generated by a lab or in the field, you will want to inspect data values and apply the criteria below, although you can apply the general Assessment Factors. Example criteria, general & specific:
  + Describe procedures for determining the quality of the existing data
  + Identify criteria or metadata for judging data or information quality, including:
  + EPA’s Five General Assessment Factors
    - Soundness
    - Applicability & Utility
    - Clarity & Completeness
    - Uncertainty & Variability
    - Evaluation & Review
  + Measurement Data
    - Accuracy
    - Precision
    - Representativeness
    - Geographic
    - Temporal
    - Technological
    - Completeness
    - Comparability
    - Age of data
    - Samples collected using the same or comparable statistical design
    - Samples or information collected over time at the same location
    - Samples represent similar matrices
    - Data generated using the same or comparable methods or SOPs
    - Data reported in comparable units of measurement
    - Data collected and managed according to an approved EPA R5-compliant QAPP or appropriate project QAPP according to the Graded Approach (PPM 13.7)

**D1-3**: **Document Data Review and Analysis**

* + Data reduction
    - Units
    - Formulas/calculations/equations
  + Data validation/accuracy
  + Statistical analyses
  + List the sources of existing data and information gathered in project QAPPs and resulting products (sources may include bibliographic databases, digital libraries or web based search engines) For websites, document versions, date and time accession stamps of any public databases queried (including number of entries searched).
  + Document the decision to accept or reject with rationale
  + Document how attainment of project objectives was based on review of existing data/information
  + Provide data/information quality considerations regarding fitness for use in any project deliverable
    - List qualifications and limitations of existing data
    - If no quality requirements exist or if the quality of the existing data cannot be determined, document with a disclaimer to indicate that the quality of the existing data is unknown.
  + If a minimum project-defined acceptance threshold for existing data cannot be met, then collect new data (see Appendix B). Each project will have its own acceptance criteria.

# DELEGATIONS

Not applicable.

# APPROVAL

Senior Information Official Date  
for Research and Development

# Appendix A. Existing Data for Addressing Research Questions & Data Limitations

| **Type of Existing Data** | **Data Source (e.g., Originating Organization, Report Title, Date)** | **Data Generator(s) (e.g., Originating Org., Data Types, Data Generation/Collection Dates)** | **How Data Will Be Used** | **Limitations on Data Use** |
| --- | --- | --- | --- | --- |
| Genomes, proteomes, transcriptomes, metabolomes, pathways, networks or any other biological molecule database used for molecular identification, classification, description, and/or molecular interaction | Any accessed external database containing biological molecule identification, description, classification, or interaction (e.g. genome, proteome, transcriptome, metabolome, pathway, network, ontology, or other molecular interaction database not produced in-house)  *Examples:* [*www.uniprot.org*](http://www.uniprot.org)*;* [*www.reactome.org*](http://www.reactome.org)*;* [*http://www.hmdb.ca*](http://www.hmdb.ca)*;* [*http://www.ingenuity.com*](http://www.ingenuity.com) | External databases provided by a vendor’s software, vendor transfer (email, website file), other website, or other external source | To identify, describe, or classify biological molecules and/or their molecular interactions | 1. Ensure QA/QC measures exist for use of any external database (assess annotation procedure and completeness) 2. Provide name (vendor software or source or website, if applicable), identify whether records are fully curated or computationally annotated, and accession date (or software vendor or version if applicable) of database   *Protein identification example: “Searches were done against a protein database with 591,237 entries (NCBI human, mouse, rat, fish and amphibian proteins from August, 2008 plus a list of 179 common contaminant proteins from Thermo Fisher).”* |
| Toxicity data | Toxicology and Applied Pharmacology.  Ninety-day toxicity study of chemical X in mice. 2005, Volume 209, pp. 600-615. | ABC Research Institute. Ninety-day toxicity study of chemical X in mice, final report. June 2003. Study sponsor: Acme Chemical Company. | To extrapolate effects observed in mice to humans, and to serve as the basis for deriving a health-based toxicity value that will be used to assess human risk. | 1. Industry-sponsored report exists in grey literature, not vetted in peer-review process.  2. Determination of the relevance of effects observed in experimental animals to humans.  2. Factors to account for uncertainty in extrapolating effects in animals to humans, and effects observed following subchronic exposure to chronic exposure situations. |
| Drinking water data | Anytown, USA:  Annual Drinking Water Check Report, 2008-2010 | TruVal Laboratories, Inc.:  Metals Drinking Water Data, Sample Collection Dates: 2/12/08, 2/1/09, 2/1/10 | To assess existing groundwater contamination | 1. Unvalidated data used to generate report 2. Limited number of wells exist to sample |
| Biological stability of materials data | ABC Database of Cellulose-Based Compounds. Empirical Research Group, Saskatchewan Canada, 1910-2009. | US EPA. The bacterial breakdown of office paper, newspaper and cardboard in soil batch experiments. Office of Solid Waste, 1972. | Model the potential for breakdown of various paper products in landfills | 1. Data generated using outdated methods and test conditions 2. Uncertainty about comparability of test conditions (soil versus municipal landfill) |
| Geographic Information System Data (a dataset that has underlying geographic encoding) | TIGER Census product, a spatial extract from the Census Bureau's MAF/TIGER database TIGER -**T**opologically **I**ntegrated **G**eographic **E**ncoding and **R**eferencing | US Census Bureau: containing features such as  Boundaries, Population Counts, Housing Unit Counts, 2010 Census Demographic Profile | To map and spatially analyze populations in proximity to industrial plants. | 1. Uncertainty in level of detail 2. Limited to 2010 census |

# Appendix B: Collecting, Screening, Evaluating, & Selecting Existing Data

Use data from several high-ranking studies

Use data from the highest ranked study

Evaluate whether new data is worth collecting

*Data Quality Evaluation*: Data quality is typically summarized according to traditional data quality indicators such as precision and accuracy, and additional considerations include the purpose/scope of original study, effectiveness of sampling design and analytical procedures, project QA/QC, and data validation and documentation completeness. Also consider general quality categories that can be distilled into discipline-specific (e.g., toxicology, epidemiology, modeling, databases, and monitoring) acceptance criteria for existing data.

*Ranking*: Studies/data are ranked according to qualitative or   
quantitative criteria derived for each category. An acceptance threshold can be defined according to scoring results to identify informative or pertinent studies/data.

*Minimum Data Requirements*: Eliminate unrepresentative or incomparable studies/data. Consider only those search results relevant to decision making. Consider screening based on data originator or publication type, research/analytical methodology, study date, study duration, effect level, or other data/study characteristics. Enhance transparency by listing basis for inclusion/exclusion of existing data.

Collection of Existing Data

Evaluation of Existing Data

Limitations of Existing Data

Decision Making

* Project-acceptance criteria
* Existing data uses

Existing data limitations

Analysis of Existing Data

Data Reduction, analysis, validation

## Figure 2: General Process

## B1. General Considerations for Existing Data (Validation, Analysis, Review)

There may be additional references that may be of use to researchers for collecting and evaluating existing data. This appendix presents general considerations for existing data with regard to data validation, data analysis and interpretation and quality, and data management and review.

**B1.1 Validation of Existing Data Sets**

Describe any data validation procedures to be applied to the existing data to ensure the use of accurate and representative project data that are fit-for-purpose in addressing the research question or questions. These procedures may include graphical or other simple, visualization, or iterative assessment approaches as well as more sophisticated statistical analysis techniques that reduce variance or minimize bias of estimates or predictors. It is important to frame the validation procedures for the questions being asked of the data.

There are modern statistical data editing approaches that aim to obtain data of sufficiently high quality for the particular research purpose rather than correcting all data in every detail as traditionally done. An important goal for statistical data editing is the detection of errors and their correction, with the main focus on errors that have a substantive influence on analysis results and interpretation. Influential data values are often outlying, although outlying influential data may be either correct or erroneous. Outlying data values are characterized as seeming not to be part of a variable’s observed distribution and can, for example, be identified using distance from the median. These deviations from the median are readily assessed graphically using boxplots. De Waal, et al. (2011) suggest standardizing deviations from the median across reporting categories using the median of their absolute values. Deviations from the median is one of several robust measures that may be appropriate in determining outlying data values that potentially warrant further assessment. Outlying data values that are also influential with respect to estimates or predictors are potential candidates for removal if considered influential errors.

Other approaches for identifying and removing outlying data values include forward search algorithms and robust tree modeling. Forward search algorithms develop model-based parameters used in calculating distance of data values from the center of the data. Only those data values within a specified threshold are retained and the process beginning with model fitting is repeated until no additional outlying data values are found. Regression tree modeling classifies data in terms of predictor variable values and partitions the original data set into groups, or nodes. Outlier-robust regression tree models down-weight data values that are relatively distal from a robust estimate of the node mean. A data value is declared an outlier when its average weight over all node splits is less than a specified threshold. Caveats on using these methods are that specifying defensible thresholds may be challenging and outlying data values may be valid. See de Waal et al. (2011) for more information on these methods and extensions to categorical and integer-valued data.

Data visualization may be a useful approach in assessing the quality of the existing data and is an important assessment tool for complicated data exploration. It is particularly well-suited for identifying bias and systematic errors in data(ASA 2013). One visualization approach for assessing the underlying accuracy and reliability of data is tableplots, which display aggregated distribution patterns of as many as a dozen or more variables(Tennekes et al., 2013). Modern approaches to combine data mining with visualization of uniformity, outlier-ness, or cluster-ness, for example, give researchers an overview and insight into data distributions (Schneiderman 2002). Visually communicating scientific data and concepts can give new insights into unexpected science Frankel and Reid 2008).

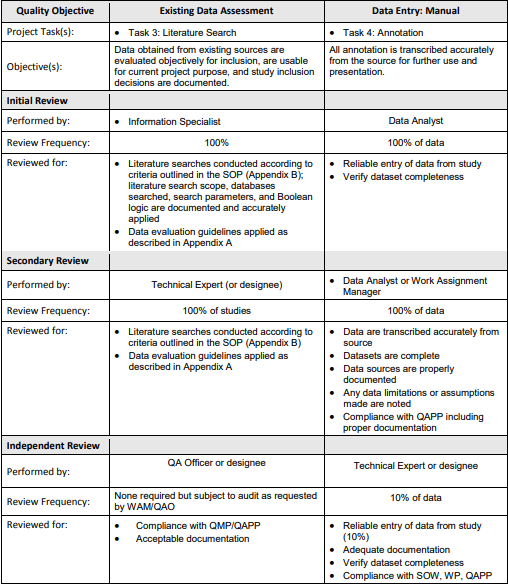
An example of a more sophisticated statistical analysis approach may include, when appropriate, cross-validation paired with modern regression methods. Ridge regression, the lasso, and other modern regression methods are characterized by reducing variance and introducing bias. Cross-validation can be used to select the tuning parameter that controls the amount of regularization in the regression in order to minimize bias. In cross-validation the full data set is randomly partitioned based on the particular research questions into, usually, 5 or 10 training sets or “folds.” Cross-validation is a useful technique when the goal is either model prediction accuracy (minimizing prediction error) or determining the right model for interpretation. A question to think about is whether the research questions are to be addressed by model prediction or model interpretation, or both. When the goal is model prediction, K-fold cross-validation uses leave-one-out training sets in choosing the tuning parameter for regression techniques that incorporate regularization, such as the tuning parameter in lasso or ridge regression. When the goal is determining the right model for interpretation, prediction error is also needed and is estimated for the model assessment. Bias correction for the minimum cross-validation estimate of the expected test error rate may be indicated with the signal-to-noise ratio is lower (Tibshirani and Tibshirani 2009). It is an important consideration that cross-validation techniques assume exchangeability and may not be appropriate if the data include structure, such as when predictor measurements are taken across time points (e.g., time-course studies).

**B1.2 Data Analysis, Interpretation, and Quality**

Discuss the data reduction procedures specific to the project, including any applicable calculations, equations, and/or computer code. Describe how the data will be summarized and/or analyzed (e.g., qualitative analysis, descriptive or inferential statistics) to meet the project objective(s). If descriptive statistics are proposed, state what tables, plots, and/or statistics (e.g., mean, median, standard error, minimum and maximum values) will be used to summarize the data. If an inferential method is proposed, indicate whether the method will be a hypothesis test, confidence interval, or confidence limit and describe how the method will be performed. Describe any additional statistical techniques to be applied, and if applicable, discuss statistical power. Describe the qualitative or quantitative assessment process to be used to generate information to determine whether the results of the data analysis are of a quality sufficient for their intended use.

When writing the descriptions for the QAPP, it is recommended to review the *Questions for Experimental, Observational or Modeling Study Design* developed by the ORD Community of Practice for Statistics (<http://intranet.ord.epa.gov/p2/nheerl/cop-statistics-study-design>).

Below, is a sample from a QAPP documenting the steps for internal quality control.



Source: An extramural QAPP from a contractor. The term WAM is no longer used but in the past it meant Work Assignment Manager. The term QAO is QA Officer.

**B1.3 Data Management and Review**

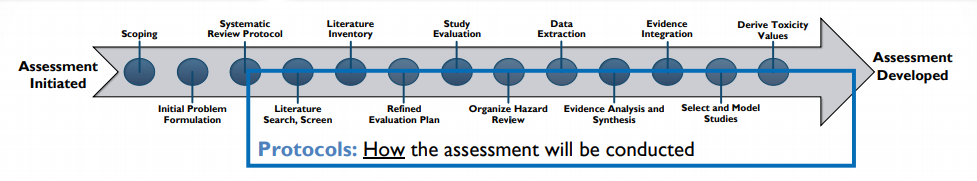
Describe data management conventions and storage requirements for both hard copy and electronic data. Describe any planned independent/external evaluation and review of the model and model design, such as scientific peer review.

## B2. Systematic Review in EPA’s IRIS Assessment Development Process

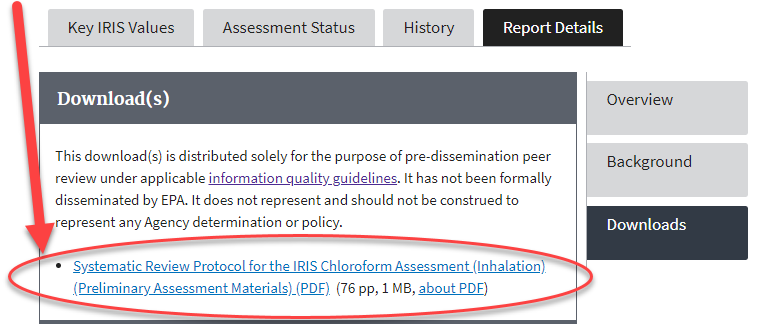
Systematic Review is a structured and documented process for a transparent literature review, (Institute of Medicine, National Academies Press, 2011). The following approach to systematic review is employed in the IRIS Program for the evaluation of epidemiological and animal toxicity studies. This is a thorough process from investigative scoping questions to evidence synthesis and integration to data content management. Materials on this topic were presented to the National Academies of Science: ***Approaches for Integrating Evidence for Chemical Assessments: EPA’s IRIS Program****, (2018),* by Dr. Kris Thayer, EPA IRIS Program Director**.**

<http://nas-sites.org/emergingscience/files/2018/02/02_04-Thayer.pdf>

Below is a figure from Dr. Thayer’s presentation, showing the process steps.



Below is the link to a document that details the systematic review protocols used for the development of the IRIS chemical assessment for Chloroform. [**https://cfpub.epa.gov/ncea/iris\_drafts/recordisplay.cfm?deid=338653**](https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=338653)



## B3. Evaluating the Quality of Individual Studies for EPA’s Integrated Scientific Assessments

Systematic review is used in the development of EPA’s Integrated Science Assessments (ISA). The ISA Preamble describes the process of searching the literature, selecting studies for consideration, evaluating study quality, synthesizing and integrating data and information (evidence) and characterizing the evidence for public health. The approach described in the Preamble is consistent with current best practices employed by others for reporting or evaluating health science data. The Preamble is used as a guide for developers of the ISAs.

See ***Preamble to the Integrated Science Assessments (November 2015), Section 4, pp. 7-8****. ISA\_PREAMBLE\_FINAL2015.PDF* [*https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310244*](https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=310244)*.*

23F

# Appendix C. General Evaluation Criteria for Existing Data vs. Measurement Data

1. **General Assessment Factors (GAF) Guidance** (2003)- The quality criteria in the GAF Guide is for evaluating existing studies- evaluating the quality and relevance of scientific and technical information used in support of an agency action and are quality considerations about the generation and documentation of information products (i.e. Assessments). The GAF presents five evaluation factors:
   1. Soundness
   2. Applicability and Utility
   3. Clarity and Completeness
   4. Uncertainty and Variability
   5. Evaluation and Review
2. **Quality criteria in the Peer Review Handbook** (2015) **and Nancy Adam’s paper,** **“*QA Reviews: How They Differ from Peer Reviews”*** (1999), differ from the GAF:
   1. Consistency
   2. Correctness
   3. Coherence
   4. Clarity
   5. Conformance
3. **DQI for Quality of Chemistry Data** (2002). Source EPA QA/G-5 **S**ection 2.1.7Quality objectives and criteria for measurement data, Table 3.

The following seven elements will be considered when assessing the quality of any chemistry data, and most existing data when applicable.

* 1. Accuracy
  2. Data precision
  3. Bias
  4. Completeness
  5. Representativeness
  6. Comparability
  7. Sensitivity

|  |  |  |  |
| --- | --- | --- | --- |
| **Side-by-Side Comparison of Published Criteria for Evaluating Data Quality** | | | |
| **Source** | **EPA General Assessment Factors (GAF) Guidance** | **EPA Peer Review Handbook, 4th Edition, Section 2.4.2;**  ***QA Reviews: How They Differ from Peer Reviews* by Nancy Adams** | **EPA DQI for Chemistry Data** (can be used for other data derived from measurements) |
| **Year Published** | 2003 | 2015; 1999 | 2002 |
| **Purpose** | Evaluate existing scientific & technical information from published and unpublished studies | Evaluate scientific and technical products (e.g. data & methods used, and conclusions) | Evaluate  measurement data |
| **Criteria Published** | 1. Soundness 2. Applicability and Utility 3. Clarity and Completeness 4. Uncertainty and Variability 5. Evaluation and Review | 1. Consistency 2. Correctness 3. Coherence 4. Clarity 5. Conformance | 1. Accuracy 2. Data precision 3. Bias 4. Completeness 5. Representativeness 6. Comparability 7. Sensitivity |

# Appendix D. SAMPLES & EXAMPLES FOR EXISTING DATA

## D1. Boilerplate QA Language for Extramural Performance Work Statements

Task #: Prepare a Quality Assurance Project Plan (QAPP)

**Sample 1**

*The contractor must submit a QAPP that addresses existing study/data collection, screening, evaluation, and analysis and ensure the existing data used are of acceptable quality and appropriate for their intended use. Existing (secondary) data is defined as the use of data that was developed for a different purpose and includes data in the published literature. The EPA General Assessment Factors Handbook provides guidance for evaluating existing data and information, (*[*https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information*](https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information)*). The QAPP should include a description of quality control checks to verify accuracy, completeness, and adherence to an established format. The QAPP shall be in conformance with EPA’s Requirements for Quality Assurance Project Plans (EPA QA/R-5). Guidance for developing QAPPs that meet EPA specifications prepared for activities conducted by or funded by EPA, are available online at* [*https://www.epa.gov/sites/production/files/2016-06/documents/r5-final\_0.pdf*](https://www.epa.gov/sites/production/files/2016-06/documents/r5-final_0.pdf) *, “*[*EPA Requirements for Quality Assurance Project Plans (QA/R-5)"*](http://www.epa.gov/quality/qs-docs/r5-final.pdf)*. The EPA QAPP ID: …… should be displayed on the QAPP cover page. Delivery Schedule: Due 5 working days after receipt of Task Order.*

**Sample 2**

*The contractor must submit a Project-Specific Quality Assurance Project Plan (QAPP) that addresses study/data collection, screening and analysis. The QAPP should include a description of quality control checks to verify accuracy, completeness, and adherence to established format. The Contractor must address in the QAPP how existing data will be considered for the Task Order. The EPA General Assessment Factors Handbook provides guidance for evaluating existing data and information, (*[*https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information*](https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information)*). Existing (secondary) data are defined as the use of data that were developed for a different purpose and include data in the published literature. The QAPP shall be in conformance with EPA’s Requirements for Quality Assurance Project Plans (EPA QA/R-5). Guidance for developing QAPPs that meet EPA specifications prepared for activities conducted by or funded by EPA, are available online at* [*https://www.epa.gov/sites/production/files/2016-06/documents/r5-final\_0.pdf*](https://www.epa.gov/sites/production/files/2016-06/documents/r5-final_0.pdf)*,“EPA Requirements for Quality Assurance Project Plans (QA/R-5)".*

*The EPA Quality Assurance manager must approve the QAPP (existing or modified) before work on this order begins. The contractor shall coordinate a meeting of the project team after completion of the QAPP and Work Plan prior to the beginning of the data analysis. Strict adherence to confidentiality and protection of personal identifiable information shall be maintained by anyone with access to the data in accordance with EPA Policy 2151.0: Privacy Policy* [*https://www.epa.gov/privacy/epa-policy-21510-privacy-policy*](https://www.epa.gov/privacy/epa-policy-21510-privacy-policy)*. The EPA QAPP ID: xxxx should be displayed on the QAPP cover page.*

*Delivery Schedule: Due 5 working days after award of Task Order.*

## D2. Sample QAPP Template

*Elements of a Quality Assurance Project Plan (QAPP) For Collecting, Identifying, and Evaluating Existing Data/Information, (2006)*

[*https://www.epa.gov/sites/production/files/2015-05/documents/assess4.pdf*](https://www.epa.gov/sites/production/files/2015-05/documents/assess4.pdf)

## D3. Example QAPPs Available Online

A selection of examples from across the agency are online and publicly available. Many are part of the ***Addendum to the Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information****.* [*https://www.epa.gov/sites/production/files/2015-05/documents/assess3.pdf*](https://www.epa.gov/sites/production/files/2015-05/documents/assess3.pdf). (December 2012) and some are on other EPA websites, see below.

*Data and Literature Evaluation for the EPA’s Study of the Potential Impacts of Hydraulic Fracturing (HF) on Drinking Water Resources, EPA/ORD/NCEA, (2014)* [*https://www.epa.gov/sites/production/files/2015-01/documents/hf\_data\_literature\_review\_qapp\_20141117\_v2\_508.pdf*](https://www.epa.gov/sites/production/files/2015-01/documents/hf_data_literature_review_qapp_20141117_v2_508.pdf)

*Supplemental Programmatic Quality Assurance Project Plan For Work Assignment 5-83 Technical Support for the Hydraulic Fracturing Drinking Water Assessment, Cadmus Group, (2013)*

[*https://www.epa.gov/sites/production/files/documents/literature-review-qapp1.pdf*](https://www.epa.gov/sites/production/files/documents/literature-review-qapp1.pdf)

*Uranium. Uravan, Colorado. QAPP for the Use of Existing Data (October 2017)*

[*https://semspub.epa.gov/work/08/100001753.pdf*](https://semspub.epa.gov/work/08/100001753.pdf)

*State of the Estuary Report NY/NJ QAPP (2009)*

[*http://www.harborestuary.org/pdf/StateOfTheEstuary2012/SOE\_QAPP\_Final\_12-03-09\_Signed.pdf*](http://www.harborestuary.org/pdf/StateOfTheEstuary2012/SOE_QAPP_Final_12-03-09_Signed.pdf)

*EPA New England QAPP for Projects Using Only Existing Data. (2009)*

[*https://www.epa.gov/sites/production/files/2015-06/documents/EPANESecondaryDataGuidance.pdf*](https://www.epa.gov/sites/production/files/2015-06/documents/EPANESecondaryDataGuidance.pdf)

## D4. Guidance and Checklists

*A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information, EPA, (2003)*

[*https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information*](https://www.epa.gov/risk/summary-general-assessment-factors-evaluating-quality-scientific-and-technical-information)

*Guidance for Evaluating and Documenting the Quality of Existing Scientific and Technical Information, EPA, (2012), Addendum to Summary of General Assessment Factors*

[*https://www.epa.gov/risk/guidance-evaluating-and-documenting-quality-existing-scientific-and-technical-information*](https://www.epa.gov/risk/guidance-evaluating-and-documenting-quality-existing-scientific-and-technical-information)

[*https://www.epa.gov/sites/production/files/2015-05/documents/assess3.pdf*](https://www.epa.gov/sites/production/files/2015-05/documents/assess3.pdf) *(Direct Link)*

*Example - Defining Assessment Factors (e.g., exclusion/inclusion criteria)*

[*https://www.epa.gov/osa/defining-assessment-factors*](https://www.epa.gov/osa/defining-assessment-factors)

*EPA Quality Staff Checklist: Using Data From Other Sources — A CHECKLIST FOR QUALITY CONCERNS, Draft,**(2000)* [*https://www.epa.gov/sites/production/files/2015-07/documents/cklist-secondary.pdf*](https://www.epa.gov/sites/production/files/2015-07/documents/cklist-secondary.pdf)

*Example Checklist in Template for Applying Assessment Factors in Collecting, Identifying and Evaluating Existing Literature, EPA ORD NCEA, (2006)*

[*https://www.epa.gov/sites/production/files/2015-05/documents/assess4.pdf*](https://www.epa.gov/sites/production/files/2015-05/documents/assess4.pdf) *(Direct Link)*

*Example Checklist for the Assessment of Existing Information (NCEA, 2006)*

[*https://www.epa.gov/osa/checklist-assessment-existing-informationsecondary-data*](https://www.epa.gov/osa/checklist-assessment-existing-informationsecondary-data)

[*https://www.epa.gov/sites/production/files/2015-05/documents/assess6.pdf*](https://www.epa.gov/sites/production/files/2015-05/documents/assess6.pdf) *(Direct Link)*

## D5. Plans, Literature Search Strategy, and Preliminary Evidence Tables

*Systematic Review Protocol for the IRIS Chloroform Assessment (Inhalation), January 2018. EPA/635/R-17/486.* [*https://cfpub.epa.gov/ncea/iris\_drafts/recordisplay.cfm?deid=338653*](https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=338653)*, click DOWNLOADS and access file CHLOROFORM\_PROTOCOL\_2018.PDF*

*Evidence Template (Chloroform), See Figure 4 on page 45 at link above.*

*IRIS Assessment Plan (IAP) for Nitrate and Nitrite (September 2017)* [*https://cfpub.epa.gov/ncea/iris\_drafts/recordisplay.cfm?deid=337470*](https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=337470) *click DOWNLOADS to access file*

## D6. Using Genomics data

[*https://www.biosharing.org/standards/reporting\_guideline*](https://www.biosharing.org/standards/reporting_guideline)

[*http://www.psidev.info/sites/default/files/MIAPE\_MSI\_1.1.pdf*](http://www.psidev.info/sites/default/files/MIAPE_MSI_1.1.pdf)

## D7. Considerations for GIS Datasets (Stream Cat)

[*https://www.epa.gov/national-aquatic-resource-surveys/nrsa*](https://www.epa.gov/national-aquatic-resource-surveys/nrsa)

# ADDITIONAL REFERENCES

12th International Conference on Grey Literature; Prague, Czech Republic 2010

Adams, NH., “Quality assurance reviews: how they differ from peer reviews”, [Qual Assur.](https://www.ncbi.nlm.nih.gov/pubmed/10386330) 1998 Apr-Jun;6 (2):75-85, 1998. EPA ORD NRMRL.<https://www.ncbi.nlm.nih.gov/pubmed/10386330>

American Statistical Association (ASA). 2013. Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society, July 2, 2013, A Working Group of the American Statistical Association, <https://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf>

Daas, P.J.H. and Ossen, S.J.L., "Metadata Quality Evaluation of Secondary Data Sources", 5th International Quality Conference, 2011. <http://www.pietdaas.nl/beta/pubs/pubs/IQConf.pdf>

De Waal T, Pannekoek J, Scholtus S (2011). Handbook of statistical data editing and imputation. Wiley, Hoboken, NJ.

Frankel F and Reid R. Big Data: Distilling meaning from data. Nature 2008 (455) 30.

Institute of Medicine, National Academies Press, 2011

Klimisch, H. J., Andreae, M., and Tillmann, U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regulatory Toxicology Pharmacology 25:1–5.

Lafornara, P., 2002 "A Graded Approach to Documenting the Use of Existing Data by Assistance Agreement Holders", Proceedings of US EPA 21st Annual Conference on Managing Environmental Quality Systems. Phoenix, AZ.

Markwiese, JT. 2003. Collection and evaluation of existing data: an ecological risk assessment perspective. Quality Assurance 10:107-116.

Shneiderman B. Inventing discovery tools: combining information visualization with data mining. Information Visualization (2002) 1, 5-12.Tennekes M, de Jonge E, and Daas PJH. Visualizing and inspecting large datasets with Tableplots. Journal of Data Science 2013 (11) 43-58.

Tibshirani RJ and Tibshirani R. A bias correction for the minimum error rate in cross-validation. The Annals of Applied Statistics, 2009 (3) 822-829

US EPA, Region 1, "Quality Assurance Project Plan Guidance for Environmental Projects using only Existing (Secondary) Data, 2009 (see <http://www.epa.gov/region1/lab/qa/pdfs/EPANESecondaryDataGuidance.pdf>

van Nederpelt, P. and Daas, P. 2012. "49 Factors that Influence the Quality of Secondary Data Sources", Statistics Netherlands, The Hague/Heerlen,. (see <http://unstats.un.org/unsd/dnss>)

Wasson, S.J., 1999. "Validating Existing Data in the Environmental Technology Verification Program", Quality Assurance, Volume 7, Pages 201-206.