

United States  
Environmental Protection  
Agency

Office of Information  
Analysis and Access  
Washington, DC 20460

February 2003  
EPA 260-B-01-006  
[www.epa.gov](http://www.epa.gov)

---

# **A Guide to Developing Secondary Information Products: Methods Review and Documentation**

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY  
WASHINGTON, D.C. 20460

OFFICE OF ENVIRONMENTAL INFORMATION

With the recent promulgation of EPA's Information Quality Guidelines (IQG), the Office of Environmental Information (OEI) is developing a series of guidance manuals, handbooks and statistical methodology papers to assist Agency programs in the analysis of their data and presentation of information in compliance with the IQG. *The Guide to Secondary use of Data in Information Products* (Guide) is the first in this series of publications. The Guide's objective is to provide Agency analysts with a conceptual framework for secondary use of existing data, i.e., the use of data that has already been collected for other purposes in a new information products.

The Guide addresses the areas that need to be considered when using existing data to develop a secondary information product. The guide does not prescribe a specific approach; however, it does advocate a process which will provide for a quality information product.

I know you will find the Guide a useful tool. We are very interested in your opinion on how comprehensive and useful the Guide is. If you have any thoughts or ideas on how to improve the document in concept, design or contents, we would appreciate hearing from you. If you have any questions or suggestions, please contact N. Phillip Ross at 202 566-0593.

Kim T. Nelson  
Assistant Administrator  
Office of Environmental Information  
United States Environmental Protection Agency

## CONTENTS

EXECUTIVE SUMMARY	3
1. Introduction	4
2. Methods Review Process	9
2.1. Pre-assessment and Planning	11
2.2. Product Development Consults	12
2.3. Finished Product Review	13
3. Documentation	15
3.1. Data Sets	17
3.2. Descriptive Statistical Analyses	20
3.3. Inferential Statistical Analyses	22
3.4. Information Graphics	24
4. Conclusion	26



## EXECUTIVE SUMMARY

The U.S. Environmental Protection Agency (EPA) collects large amounts of data to manage regulatory programs, set regulatory targets, and take compliance action. With the emergence of “right to know” programs and widespread access to the Internet, the audience for EPA’s environmental data has expanded enormously. EPA data are being used more widely and in a broader variety of applications than ever before. Often the data are used for purposes that were not anticipated when the data were collected. These “secondary information products” are the subject of this guidance document.

The Office of Information Analysis and Access in EPA’s Office of Environmental Information has developed this document to assist EPA staff in developing new information products and supporting the public’s right to know about environmental conditions. It is not meant to be a set of standards that all are required to follow. It is intended to help EPA analysts develop secondary information products in a rigorous and credible manner. The guidelines articulated in this document are especially pertinent to EPA staff, partners, and contractors working to develop integrated information products.

The guidance consists of two parts:

- The Methods Review Process outlines an approach to assure that secondary information products are planned and developed using appropriate analytical and statistical tools, and carefully reviewed under rigorous but flexible review procedures. Methods review should be considered (1) in the pre-assessment and planning phase of a project, (2) during product development, and (3) post-product development.
- Easy-to-follow Documentation Guidelines are provided for four types of secondary information products: data sets, descriptive statistical analyses, inferential statistical analyses, and information graphics. Documentation allows interested third parties to review how a statistical or analytical product was developed, and if desired, replicate the outcome or findings.

The procedures in this document are intended to support and complement existing Agency processes, such as Quality Assurance, Information Quality Guidelines (to be promulgated Oct. 1, 2002), and Peer Review. While the guidelines are not mandatory, they are highly recommended as a way of ensuring that EPA’s secondary information products maintain the highest standards of quality, credibility, and reproducibility on which the Agency’s reputation rests.

## 1. INTRODUCTION

EPA's Office of Environmental Information (OEI) was created to improve the way EPA collects, manages, integrates, and provides access to environmental information. One of the primary functions of OEI is to provide the public with high-quality and useful information on environmental quality, status, and trends. In many cases, the new information products are derived from EPA's primary data systems (e.g., Toxics Release Inventory).

Many of EPA's primary databases were originally compiled to track the requirements of particular regulations or statutes. Increasingly, EPA data are being used for purposes not anticipated when they were originally collected. This tendency has accelerated due to the ease of access to data afforded by the Internet and the proliferation of easy-to-use software applications and statistical packages. The results of these applications are called secondary information products.

All of the information EPA publishes or makes available must be of high quality and analytically rigorous. This applies as much to secondary information products as to primary data collections. However, while primary data collection is subject to rigorous peer review and quality assurance protocols<sup>1</sup>, secondary information products are often developed outside the purview of routine program activity and tend not to be subject to formal validation or review procedures. As OEI and other EPA offices develop new information products, there is a need for guidance to help assure the credibility and rigor of the Agency's information inventory. This document provides guidelines for developing credible secondary information products, especially those disseminated to the public.

*As OEI and other EPA offices develop new information products, there is a need for guidance to help assure the credibility and rigor of the Agency's information inventory.*

---

<sup>1</sup> Guidance documents are currently available on planning and analysis for data collected under the Agency Quality System. See: [http://www.epa.gov/quality/qa\\_docs.html](http://www.epa.gov/quality/qa_docs.html)

## *Secondary Information Products*

A joint EPA/State Action Team recently defined “significant information products” as follows:

A product under development or major modification by EPA, which derives from Federal, State, local, Tribal, and/or other organizations' data, and a State product that is regional or national in scope and aggregates data from more than one State.<sup>2</sup>

As suggested by this definition, most EPA data are initially collected or generated by first-order data providers, usually industry or other regulated entities and states. Data are then transferred from the original provider to EPA and/or state data systems, where they are maintained and used to support various regulatory and program administration activities. Increasingly, these data are being used or adapted for new, possibly unanticipated, “secondary” applications.

Secondary applications include integration, consolidation, graphical representation, transformation, aggregation or disaggregation, analysis, and context-specific uses of original program data. Secondary information products include both paper-copy and electronic versions. Examples of recent OEI secondary information products are shown in Exhibit 1.

### **Exhibit 1 – Examples of Secondary Information Products**

- The Envirofacts Warehouse contains extracts from EPA databases on air quality, chemicals, facility information, grants/funding, hazardous waste, risk management plans, Superfund, toxic releases, water permits, drinking water, drinking water contaminant occurrence, and drinking water microbial and disinfection by-product information.
- The EnviroMapper application allows users to view spatial data at the national, state, and county levels, as well as utilize GIS functionality, such as displaying multiple spatial layers, zooming, panning, identifying features, and querying single points.
- Toxics Release Inventory (TRI) Public Data Release Report provides annual summaries, analyses, and comparisons of TRI data by year. It contains detailed analyses and supporting tables for releases and other waste management of TRI chemicals.
- TRI Explorer is a downloadable software application that allows users to explore the EPA Toxic Release Inventory data to learn about the presence of toxic chemicals in their communities and the amounts of releases of these chemicals into the environment.

<sup>2</sup> <http://www.epa.gov/ipbpages/significant.htm>

Four broad categories of secondary information products are likely to be developed by OEI and other EPA offices working to satisfy Goal 7 of the EPA Strategic Plan.<sup>3</sup> These categories are not necessarily independent of one another.

1. Data sets are extracts of data from systems that house primary data with no accompanying analyses or graphics. Sometimes, the data sets involve additional transformations or “cleaning up” of the primary data.

2. Descriptive statistical analyses deal with the characteristics of the data. They include simple univariate measures of dispersion and central tendency such as means, medians, standard deviations, and percentiles of the variables. These analyses can also produce bivariate information describing the relationship between variables, such as regression prediction equations or correlation coefficients.

3. Inferential statistical analyses involve advanced statistical methodology and are concerned with extrapolating information beyond the immediately available data. Inferential statistics can include hypothesis testing, confidence intervals, and other types of analyses intended to characterize a population of interest based on sample data.

4. Information graphics include visual displays and presentations of data in the form of bar charts, trend lines, and other graphic representations, as well as maps created through geographic information systems (GIS) and other spatial representations using cartographic/geographic frames.

---

<sup>3</sup> The public and decision makers at all levels will have access to information about environmental conditions and human health to inform decision making and help assess the general environmental health of communities. The public will also have access to educational services and information services and tools that provide for the reliable and secure exchange of quality environmental information.

***The goal is to raise the awareness of EPA staff on the importance of asking the right questions early on in a project, and following up with careful documentation and adequate review.***

## *Purpose of This Document*

The purpose of this document is to give EPA staff voluntary guidance on developing reliable, well documented secondary information products, especially those that are intended to be disseminated to the public.

EPA already has in place a Quality System<sup>4</sup> designed to ensure that data used to support Agency actions and decisions are collected or generated under rigorous Quality Assurance (QA) and Quality Control (QC) procedures. (See Exhibit 2.) This document is not intended to duplicate the Agency-wide Quality System, nor to provide a how-to manual for developing data products. The goal instead is to raise the awareness of EPA staff to the importance of asking the right questions early on in a project, and following up with careful documentation and adequate review.

### Exhibit 2 – The EPA Quality System

Established under EPA Order 5360, the Agency-wide Quality System is a management framework that provides elements to plan, implement, document, and assess the effectiveness of Quality Assurance and Quality Control activities applied to environmental programs conducted by or for the Environmental Protection Agency. The Quality System includes the following functions:

- Establishing quality management policies and guidelines for the development of organization- and project-specific quality plans;
- Establishing criteria and guidelines for planning, implementing, documenting, and assessing activities to obtain sufficient and adequate data quality;
- Performing management and technical assessments to ascertain the effectiveness of QA/QC implementation; and
- Identifying and developing training programs related to QA/QC implementation.

Quality management systems must conform to ANSI/ASQC specifications and include the following:

- An individual assigned to serve as a Quality Assurance Manager (QAM), who functions independently of data collection/generation activities;
- A Quality Management Plan (QMP) that documents the organization's quality policy, identifies the specific programmatic activities subject to the Quality System;
- Annual assessments of the effectiveness of the quality system; and
- Approved Quality Assurance Project Plans (QAPPs) for all applicable projects and activities.

---

<sup>4</sup> [www.epa.gov/quality/](http://www.epa.gov/quality/)

Essentially, this guidance seeks to advocate the use of the following principles for handling data:

- Be aware of the pitfalls of using data that have been collected in different ways, at different levels of accuracy, etc.
- Be clear about what you want to accomplish with the data, and why.
- Carefully document what you do with the data, from the raw information to the finished product, including assumptions you make along the way, additions, deletions, and transformations to the data (including “cleaning up” the data), and software used.
- Rather than waiting until the end of the project for a final review – when it can be costly to fix major problems – put a team in place at the start to help review the plans, deal with implementation issues, and determine what review is needed for the finished product.

The remainder of this guide is divided into two parts:

1. Methods Review Process outlines key considerations for ensuring adequate review of an information product by internal and/or external reviewers, during the course of the project.
2. Easy-to-follow Documentation Guidelines are presented for the four categories of secondary information products.

## 2. METHODS REVIEW PROCESS

The methods review process recommended here is a voluntary, flexible approach that can be customized to meet the requirements of specific information products. The process utilizes expert input from multiple disciplines (as needed) throughout the product development life cycle, thereby avoiding protracted review and correction toward the end of the project. By addressing potential problems before they occur, the methods review process should 1) reduce the total time and resources expended during product development; 2) enhance product quality and credibility; and 3) expedite product development and public dissemination.

The purpose of the methods review process is to promote credibility of the information product by ensuring the following:

- Quantitative feasibility: The issue being described by the product lends itself to quantitative description, the product is the appropriate quantitative measure, and it is feasible to develop the product.
- Appropriate use of statistical methods: Once the feasibility of the project is determined, the statistical tests or methodologies are chosen in light of the issue examined and the data used.
- Correct implementation of methodology: In addition to appropriate use of statistical methods, the underlying assumptions are appropriate and the methodology is implemented without error.
- Product documentation: Details about product development, including information about data and methods, are documented and provided, allowing reproducibility of the product.

*...ensure that the information product will be developed using analytical methods that are both feasible and appropriate given available data and the purposes of the project.*

- **Product understandability:** The product is made available in a clear and understandable manner. Graphic and/or narrative representations are consistent with mathematical/statistical findings.

Elements of the review process are described below.

### ***Putting Together a Methods Review Team***

The methods review process described here is based on the time-honored concept that rigorous review is best conducted by technically competent individuals who are not directly involved in the project itself. The product developers, of course, are responsible for the conceptualization, generation, and completion of the information product. We recommend that the product developers select a “methods review team” that can provide expert statistical and scientific review for the project. The methods review team should provide consultation and review of statistical and scientific methods used in the development of information products. The review team should be comprised of multidisciplinary reviewers with expertise in relevant areas including statistics, cartography, quantitative geography, computer programming, operations research, etc. Once again, it is important that reviewers have little involvement with the generation of the information product.

### ***Three Phases of Methods Review***

In general, the methods review team is expected to review a project to ensure credibility in terms of the five components mentioned above. The diverse nature of information products will require the review process to be tailored to meet the needs of each project. In some instances, a project may require only a minimal methods review, while other projects may involve sophisticated analytical designs and/or innovative visualization techniques, which require significant review. It is appropriate to consider undertaking methods review at three phases in a project:

1. **Pre-assessment and Planning:** It is recommended that product developers meet early on with statisticians and other appropriate disciplinary or subject-matter experts to assess the feasibility and appropriateness of methods proposed to be used in developing the product, and to discuss needs for peer review and technical evaluation.
2. **Product Development Consults:** As product development proceeds, statistical and other information science experts should be invited to consult with product developers on various aspects of the information product as required.
3. **Finished Product Review:** Product review activities could range from in-house evaluations to full-scale, external peer reviews, such as required under ORD guidelines.

These three phases are elaborated in Sections 2.1, 2.2, and 2.3 below.

## 2.1 Pre-assessment and Planning

During the conceptual stage of information product development, it is recommended that developers select a review panel comprised of statisticians and experts from appropriate disciplines. Product developers should meet with the methods review team to ensure that the information product will be developed using analytical methods that are both feasible and appropriate given available data and the purposes of the project.

During this pre-assessment step, questions such as the following should be posed and answered:

- Is the concept of the information product feasible from a quantitative perspective?
- Which statistical methods are appropriate for developing quantitative characteristics?
- What other information sciences will be used – such as cartography, quantitative geography, computer programming, or operations research?
- How will quantitative methods be implemented in terms of hardware and personnel resources?
- How will the information be presented?

The pre-assessment meeting should scope out a methods review plan that indicates the types of reviews needed at particular milestones during the course of the project, and develop a preliminary schedule for those reviews. The types of reviews needed will depend on the complexity and scope of the information product. For example, a methods review plan could include any or all of the following review mechanisms:

- Keeping statistical and technical reviewers apprised of project progress through regular e-mail, memo, and report distribution lists.
- Meeting with review team members at major milestones, such as initial analysis of data, initial development of statistical graphics templates, completion of questionnaire design, development of sampling plan, etc.
- Determining the criteria for fulfilling documentation requirements in terms of topics or issues to be included, and the depth of answers to be sought.
- Initiating an extramural peer review process for the information product.

Pre-assessments should be timely and should be integrated into the project schedule. Ideally, they should provide the product developers with assurance regarding quantitative methods and product time lines.

## 2.2 Product Development Consults

Methods review during product development should consist of regular and iterative consultations between product developers and review panelists providing expertise in the areas of statistics and other information sciences. The models and methods used in the project should be evaluated and critiqued with regard to their purpose, major defining and limiting considerations, theoretical basis, parameter estimation, data quality and quantity, key assumptions, performance measures, documentation and users guide, and retrospective examination.

Examples of consultations include:

- Discussing the limitations of a data set and the implications for employing specific quantitative methods.
- Reviewing the sampling plan for a source data set.
- Providing statistical expertise in evaluating the usefulness of complex databases from other organizations or federal agencies.
- Reviewing the use of maps in terms of conventional cartographic standards.
- Consulting on statistical graphics and advising on new and useful ways of conveying statistical information.

As part of the process, the project staff should document the information product as it is being developed, thus ensuring that important decisions made and transformations of the data are not forgotten or lost. Such ongoing documentation will also assist reviewers in assessing the quantitative methodology being used. (See Section 3 below for documentation checklists.)

## 2.3 Finished Product Review

A review of the finished information product is critical to ensuring its credibility. Unlike the pre-assessments and methods review during development, the finished product review is not generally a collaborative effort. Instead, it is an opportunity to have a product objectively reviewed by external experts for content and presentation prior to its release to the public.

As noted earlier, finished product reviews can range from in-house evaluations to full-scale external peer reviews. The project developers, working in concert with the methods review team, should determine what level of final review is needed. A full peer review will not always be required for all secondary information products; its need should be determined on a case-by-case basis. It is recommended that Agency offices adhere to EPA's Science Policy Council (SPC) Handbook on Peer Review<sup>5</sup>, which offers criteria for undertaking a full peer review, and outlines the following steps in conducting a peer review:

1. Determining the key issues to be addressed: The SPC Handbook points out the importance of formulating a "clear, focused charge that identifies recognized issues and invites comments or assistance..." The charge to peer reviewers usually makes two general requests. First, it focuses the review by presenting specific questions and concerns that EPA expects the reviewers to address. Secondly, it invites general comments on the entire work product. The questions should be specific enough to elicit helpful comments, but not so specific that they preclude creative responses (unless very specific points need to be addressed).

*...an opportunity to have a product objectively reviewed by external experts for content and presentation prior to its release to the public.*

---

<sup>5</sup> Peer Review Handbook, EPA 100-B-98-001, January, 1998. (<http://www.epa.gov/ORD/spc/sopmenu.htm>)

**2. Setting the schedule and format:** The schedule should include sufficient time for recruiting appropriate experts, allowing for the review and comment period, incorporating peer review comments in the product, obtaining any necessary clarifications, and finalizing the information product prior to publication. Options for peer review format include mail review, face-to-face meetings, and one-time or multiple meetings.

**3. Selecting peer reviewers:** Peer reviewers can be selected from inside or outside EPA, but in order to ensure their objectivity, they should have little or no involvement with the actual generation of the information product. The key criteria should be their technical expertise, availability of time, and independence from the office producing the product, as well as the balance of viewpoints across the panel. A peer review panel may include just a few individuals or up to ten or more, depending on the issues involved and the range of expertise needed. As the SPC Handbook notes, “a review conducted by one individual will rarely provide the depth of commentary required to improve the work product... [or] the range of views and richness necessary to ensure improvement.”

In general, we do not recommend keeping the reviewers anonymous to the project staff. However, for highly charged and/or complex projects, anonymity can be helpful. If important policy issues are involved in the project in addition to qualitative analyses, the review team should include reviewers with expertise in both areas, or consider scheduling sequential reviews. For example, the peer review for EPA’s Center for Environmental Information and Statistics’ Customer Telephone Survey involved reviewers with knowledge of necessary quantitative skills as well as reviewers for policy issues.

Reviewers should be asked if they have any real or perceived conflicts of interest, based on current and prior work and clients. This information, as well as the selection of reviewers, should be documented.

**4. Completing a peer review:** Materials to be provided to peer reviewers include the documentation of the information product, tools used to generate the product (such as a questionnaire survey form), relevant written memos and other background materials related to the information product, and the information product itself. The final phase of the peer review involves evaluating and transmitting comments from the peer reviewers to the project team.

The goal of the methods review process described in this section is to bring to bear adequate planning and review in the development of secondary information products in order to strengthen their integrity and ensure their credibility. The following section provides a series of checklists suitable for documenting different types of secondary information products. Documentation, too, is an essential part of a high quality, credible, and reproducible information program.

### 3. DOCUMENTATION

EPA’s information products need to be well documented in order to adequately serve a broad user community. Analysts require detailed documentation to fully understand the steps taken and methods employed in product development. Such information can be used to help replicate results.

As introduced in Section 1, this document is structured in terms of four broad types of secondary information products: data sets, descriptive statistical analyses, inferential statistical analyses, and information graphics. EPA information projects tend to involve some or all of these different types, in various combinations. Data sets are developed and disseminated for other users to conduct their own analyses. Data sets are also created for the explicit purpose of performing descriptive, inferential, and graphical analyses by Agency staff. Exhibit 3, below, lists the recommended areas of documentation for each of the four basic types of information products.

As depicted in Exhibit 3, data sets are the foundation of all analyses and their documentation should be viewed as a basis for complete documentation of all information products. Additional areas of documentation, shown in the bottom half of Exhibit 3, may be needed for the unique or special requirements posed by descriptive statistics, inferential statistics, and information graphics.

<b>Exhibit 3 – Areas of Documentation by Information Product Type</b>					
Documentation Areas		Data set	Descriptive Statistics	Inferential Statistics	Information Graphics
Core Documentation Areas	Data sources				
	Concepts and coverage of data set				
	Descriptions of variables				
	Data standards				
	Data manipulations and caveats				
	Data quality and timeliness				
	Data management and output preparation				
Areas of Special Emphasis	Need for and type of descriptive statistics				
	Descriptive analysis algorithms				
	Descriptive analysis methodology				
	Need for inferential statistics and null hypothesis				
	Sampling issues				
	Inferential analysis algorithms				
	Inferential analysis methodology				
	Need for information graphics				
	Generation of graphics				
	Algorithms for graphics				

The remainder of this section includes checklists to assist EPA staff in documenting information products. The checklists enumerate the kinds of information required for thorough documentation of each information product category. Phrased in a “question” format, analysts are encouraged to review potential products in terms of all relevant criteria. These checklists are only guidelines and not intended to be comprehensive or minimum requirements; staff working on a given information product are in the best position to know what additional information is relevant for proper documentation of that product.

Some questions from the checklists can be answered by reviewing metadata associated with the data used to develop secondary information products. For example, most monitoring data have metadata associated with them because they are subject to EPA’s Quality System. Details on the information available from data sets subject to EPA’s Quality System are found in EPA Order 5360.1 Policy and Program Requirements for the Mandatory Agency-wide Quality System, EPA Manual 5360; EPA Quality Manual for Environmental Programs; and American National Standard: Specifications and Guidelines for Quality Systems for Environmental Data Collection and Environmental Technology Programs, published by the American Society for Quality Control.

### 3.1 Data Sets

Documentation for a data set helps the user determine what the data represent, what their limitations are, and how to use them appropriately. Data are often “cleaned up” before they are released. Therefore, it is important to document the source(s) of the data as well as any manipulations or special characteristics to provide the user with all available background information on the data source or collection methods necessary to interpret the data. Adequate documentation of a data set typically includes information on data sources, concepts and coverage of data, description of variables, data standards, data manipulations and caveats, data quality and timeliness, and data management and output preparation. A sample checklist in the form of questions is provided for each of these categories.

#### *Data Sources:*

- Which offices or organizations were involved in the collection, compilation, and publication of the data?
- In which key publications can the source data be found, including electronic media?
- Which articles on methodology or methodological descriptions were used or are relevant? (Provide citations.)
- In which publications can the source data be found? (Provide citations.)
- What is the historical time period for which the source data are published or available?
- Does the source data include information of a confidential nature? Have the conditions of confidentiality been clearly outlined?

#### *Concepts and coverage of the data set:*

- What is the purpose of collecting/publishing the data set?
- Who is the intended audience for the data set?
- What are the general characteristics of the population included in the data set?
  - Are the data from a rare population, a clustered population, etc.?
- Do the data have temporal characteristics?
  - What is the run of the data? Have the data been de-trended in any way?
- Do the data have spatial characteristics?

- What is the geographical coverage of the data?
  - At what level were the data collected: national? regional? state? county? Census tract?

***Description of variables:***

- What variables are used and what are they intended to measure?
- In what units of measurement are the variables expressed?
- Which variables are qualitative and which are quantitative?
  - If categorical variables are present, how are they coded and what was the original qualitative description (e.g., male = 0, female = 1)?
  - Is this coding universal for all data collected?
  - How is the coding procedure chosen?
- Are the variables clearly explained for a user in the target audience?

***Data standards:***

- Are there recognized international/national standards for compiling the data? If so, have they been used?
- Are there standard classifications used to describe the data?
- Is there international comparability with the data collected?
  - Are there any departures from international standards, such as differences in units or definitions of variables?

***Data manipulations and caveats:***

- Have the data been manipulated?
  - Are there outliers? If so, how are they identified and dealt with?
  - Are the data normalized?
  - Are there multiple responses at a sampling location? If so, how are they dealt with?
  - How are non-responses in the data handled? Are they imputed? How and why is a particular method of imputation used?
  - Are the data generated from a statistical model? What data are used to generate this model? How was the model chosen?
  - What level of data is derived from model-generated observations (e.g., local levels interpolated from state or national data)?
  - Are reliability/cross-validation tests used to test the validity of this model for the data?
- Are there any caveats about the data that would suggest limitations?
  - Are there features of the data that might cause mistakes in efforts to reproduce the data from the raw observations if a user was unaware of them?
  - What disclaimers are relevant to the data?

### ***Data quality and timeliness***

- How are error sources dealt with when extracting the analysis data set?
- Are there any breaks in a time series?
  - Are there any changes to definitions, collection methods, new weighting systems, etc., which could influence data comparability or the coherence of the time series?
- What is the policy for correcting errors found after the data have been collected (e.g., transcription errors, reporting errors)?
- Are data being compared against published observations of the same variables, if such a publication exists?

### ***Output preparation and data set management***

- How is the data set stored?
- Which person or office is responsible for updating the data set?
- What are the access rules to the data set?
- Can the data be downloaded?
  - Do instructions on downloading accompany the data set?
  - What formats can the data be downloaded into?
- If the data set is intended as input for other information products, are the data transformed in any way (e.g., to achieve approximate normality)?
- Will data from different sources be merged or analytically integrated? If so, have data inter-relationships been made clear?
- What assumptions are made about the data for further analyses?
  - Is normality assumed?

### 3.2 Descriptive Statistical Analyses

Documentation of descriptive statistical analyses should first include information about the data set used in the analysis (see Section 3.1). Additional documentation includes: the need for and type of descriptive statistics, the algorithms used to generate the statistics, and sufficient details about the methodology. A checklist of questions to be answered in documenting descriptive statistical analyses is provided here.

#### *Need for and type of descriptive statistics*

- What is the purpose of calculating these descriptive statistics?
- Who is the intended audience?
- Which specific descriptive statistics are calculated and on which variables?
- Which statistical software packages are used?

#### *Descriptive analysis methodology*

- What type of methodology is used?
- What are the methodological assumptions?
- How are errors in the data corrected? Are there any corrections for bias?
- What are the limitations of the methodology?
- What are the limitations associated with the output?
- Are there any special considerations in interpreting the statistics?

#### **Descriptive Analyses: National Air Quality and Emissions Report**

An example of documentation of descriptive analysis is provided in the 1997 National Air Quality and Emissions Trends Report. This report tracks concentrations of air pollutants as measured by air quality monitors located in and around urban areas and other locations in the country. The appendix includes documentation of the analysis. Documentation of methodology includes:

- Details on the sources of air quality data, including monitoring sites selection criteria, details on monitor operation and maps indicating monitor locations;
- Protocols used in the development of air quality trend statistics; and
- A discussion of the emission estimation methodologies and models used.



[www.epa.gov/oar/aqtrends.html](http://www.epa.gov/oar/aqtrends.html)

### *Descriptive analysis algorithms*

- Are algorithms generated and used to calculate any statistical results? (Note: this includes computer interfaces on Web sites used to calculate statistics.)
- What are the algorithms generated and what are their purposes?
- What language is used to generate the algorithms?
- Have the algorithms used been tested to verify the accuracy of the output?
  - Is a test data set used to verify the output of the algorithm?

### 3.3 Inferential Statistical Analyses

Documentation is particularly important for inferential analyses due to the in-depth nature of the statistics used in such analyses. It is recommended that the data set be described along with the methods used to manipulate or transform the data (see checklist in Section 3.1.) Additionally, the need for the analysis, the null hypothesis, sampling issues, methodology and algorithms should also be documented. Below is a checklist to assist in the documentation of inferential statistical analyses. The questions are not intended to be all-inclusive, but will help direct the documentation process.

#### *Need for inferential statistics and null hypothesis*

- What are the research hypotheses driving this analysis?
  - What questions are being answered with this analysis?
  - What is the purpose of calculating these inferential statistics?
  - Who is the intended audience?
- What is the null hypothesis for this analysis?
  - What is held to be true regarding the issue under question?

#### *Inferential analysis methodology*

- What type of inferential analysis was conducted?
- What statistical software packages were used?
- How were significant results identified?
- Were goodness-of-fit tests conducted? If so, what methods were used?
- What tests were chosen to show what?
  - What variables were used in these tests?
  - Were the assumptions for these tests met?  
How were assumptions tested?

#### **Inferential Analyses: Technical Support Document for Hazardous Waste Combustion MACT Standards**

An example of good documentation of inferential analyses is provided by EPA's Emission Estimates and Engineering Costs Technical Support Document, which describes the costs of compliance with the Maximum Achievable Control Technology (MACT) standards for controlling emissions from hazardous waste combustors (HWCs). Chapter 2 of the document explains the approach used to develop national emissions estimates for HWCs. Details are provided on methodological issues, such as:

- Data selection criteria and limitations;
- The "imputation" methods used for estimating emissions;
- Treatment of non-detected data; and
- Stack height/diameter and other parameter estimation.

HWC MACT Webpage

See Background Documents page



[www.epa.gov/epaoswer/hazwaste/com bust/](http://www.epa.gov/epaoswer/hazwaste/com bust/)

- Was cross-validation of the results performed? If so, what methods were used?

### *Sampling issues*

- What were the sampling units (i.e., what was sampled)?
- What was the sample size and how was it chosen?
  - Was a sampling method used to determine an optimal sample size?
  - What were the constraints in choosing this sample size?
- What sampling method was used?
  - Was a particular sampling scheme chosen? If so, how was it chosen?
  - Were the characteristics of the population considered in the choice of sample scheme?
  - Was inferential information on the estimators a factor in choosing a particular sampling methodology?
  - Was an effort made to minimize the variance of the estimator?
- If a survey was used, how was the questionnaire delivered and filled out (e.g., telephone, mail, personal interview)?
  - Provide a copy of the questionnaire.
- What was the non-response rate?
- How were sampling errors and other errors corrected? Were there any corrections for bias?
- Are there any breaks in a time series?
  - Are there any changes to definitions, collection methods, new weighting systems, etc., which could have had an influence on data comparability or the coherence of the time series?
- What is the revision policy regarding the data?
- Have the data been compared against published observations of the same variables, if such a publication exists?

### *Inferential analysis algorithms*

- Are algorithms generated and used to calculate any statistical results? (Note: this includes computer interfaces on Web sites used to calculate statistics.)
- What are the algorithms generated and what are their purposes?
- What language is used to generate the algorithms?
- Have the algorithms used been tested to verify the accuracy of the output? Is a test data set used to verify the output of the algorithm?

### 3.4 Information Graphics

An information product that consists only of information graphics, or an interface that produces these graphics, still requires documentation. Both the underlying data set (see checklist in Section 3.1.) and the details of the information graphics should be documented. The details include the need for the graphics, special data considerations, and the actual generation of the graphics. Typical questions to be answered are listed below.

#### *Need for information graphics*

- What is the purpose of generating this information graphic?
- Who is the intended audience?

#### *Generating the graphics*

- Does a data set accompany the graphics?
- Is the graphic a result of an interface using Internet media?
- What software package is used to generate these graphics?
- What map projection is used?
- What is the time period of the data used in the graphic?
- How are the data divided?
  - o What are the natural data breaks, quartiles, ranges, etc.?
  - o What are the numbers in the classes?
- What is the geographical coverage of the data?
  - o Specify the level at which the data were collected: National? State? Regional?
  - o What is the scale of the map data layers? (e.g. 1:24,000)
  - o What level of information is used in the graphic?

#### **Information Graphics: EnviroMapper**

An example of documentation of maps can be found in EPA's EnviroMapper application, which provides users with interactive Geographic Information System (GIS) capabilities using spatial data for the United States.

Documentation for EnviroMapper includes:

- Data sources;
- Metadata;
- Map scale; and
- Significant spatial features.



[www.epa.gov/enviro/html/mod/](http://www.epa.gov/enviro/html/mod/)

### *Algorithms for graphics*

- Are algorithms created to generate these graphics? (Note: this includes computer interfaces on Web sites used to produce graphics.)
- What are the algorithms generated and what are their purposes?
- What language is used to generate the algorithms?
- Have the algorithms used been tested to verify the accuracy of the output?
  - Is a test data set used to verify the output of the algorithm?

#### 4. CONCLUSION

The methods review process and documentation guidelines outlined in this document are intended as a “preventative” approach – to uncover technical problems and unresolved issues before they become obstacles to the publication and use of environmental information products. The goal is to avoid lengthy, protracted review of a product once it is finished, by involving experts in a project’s early development and by continuing the collaborative involvement of experts during the product’s development. Such an approach will improve overall product quality and ultimately reduce the time and resources needed to develop products.

Documentation is a record of what has been done to generate an information product. While it is useful to note issues and limitations related to the data, documentation should primarily illustrate the steps taken to generate the information product. Thus, the documentation should serve as a formula that can, if necessary, be used to replicate the information product. Full documentation allows other researchers to understand our data, evaluate our methods, and reproduce our results. All of these are essential components of the scientific process. Good documentation procedures lend credibility to EPA’s work, and ensure that EPA remains a reliable source of high-quality environmental information and analysis.

For more information on this document, assistance in implementing a methods review process, and/or developing documentation, please contact:

N. Phillip Ross  
Office of Information Analysis and Access  
Office of Environmental Information  
(202) 566-0593  
ross.np@epa.gov