United States
Environmental
Protection Agency

Science Advisory
Board
Washington DC

EPA-SAB-EC-99-010
February 1999
www.epa.gov/sab

AN SAB REPORT: DATA SUITABILITY ASSESSMENT

REVIEW OF THE CENTER FOR ENVIRONMENTAL INFORMATION AND STATISTIC'S (CEIS) DRAFT DATA SUITABILITY ASSESSMENT OF MAJOR EPA DATABASES

February 19, 1999

EPA-SAB-EC-99-010

Honorable Carol M. Browner
Administrator
U.S. Environmental Protection Agency
401 M Street, SW
Washington, DC 20460

Subject: Secondary Data Use Subcommittee Review of CEIS's Draft "Data Suitability Review of Major EPA Databases"

Dear Ms. Browner:

The Science Advisory Board (SAB) reviewed the draft Data Suitability Assessment dated November 11, 1998 prepared by the Center for Environmental Information and Statistics (CEIS). This is a part of the ongoing effort by the SAB to help EPA meet the challenges it faces from release of Agency databases for secondary use. The review was done by the Secondary Data Use Subcommittee ("SDUS" or "the Subcommittee) of the SAB's Executive Committee at its meeting on December 15, 1998.

The Subcommittee found that the Suitability Assessment was carefully and thoughtfully assembled, and that it was an excellent first step in a process that should lead to a more widespread and productive utilization of historic EPA datasets. The Subcommittee members agreed that what had been included in the draft CEIS document is appropriate for evaluating the general suitability of databases for a range of secondary uses. There was also a consensus that additions to the present draft would improve the usefulness of the data bases to secondary users. Several kinds of additions are suggested, specifically:

a) Qualitative additions to the present review of the databases (e.g., a glossary).

b) More quantitative additions that would be appropriate during the second, more quantitative stage of review, which CEIS plans for the future (e.g., quantitative precision information).

c) Activities or documents in addition to the review of the data bases (e.g., provision for user feedback).

The Subcommittee did not attempt to prioritize the recommendations contained in its report, but focused on providing a range of ideas and recommendations. This approach seemed likely to be more helpful to the CEIS at this point. However, the Subcommittee recognizes that priorities might be useful to the Agency as it decides how to use limited resources in the immediate future.

The Subcommittee would be happy to consider providing additional discussions to address priorities, if that would be helpful, while, at the same time, recognizing that factors other than scientific considerations will influence the Agency 's final decisions on priorities.

In addition to the review of this document, the SDUS met on December 16 for a briefing on the CEIS approach to the analysis of geographically based environmental indices and for a broader discussion with Associate Deputy Administrator Margaret Schneider and other Agency personnel on information management at EPA. The SDUS also discussed the need for and possible scope of future Subcommittee activities and responses to Agency requests for scientific advice. The Subcommittee was pleased to have the opportunity to learn more about the Agency's initiatives on information management and the high priority that they have received.

The SAB congratulates the Agency on being at the forefront of a new focus on information, and for conducting the suitability assessment reviews, which will benefit users outside and inside the Agency. Based upon the productive interactions at the December meeting , the SAB looks forward to future meetings with Agency officials that will assist EPA in achieving optimal information management decisions. We look forward to the response of the Assistant Administrator for the Office of Policy to the advice contained in this report.

Sincerely,

/signed/
Dr. Morton Lippmann, Chair
Secondary Data Use Subcommittee
Science Advisory Board

/signed/
Dr. Joan Daisey, Chair
Science Advisory Board

NOTICE

This report has been written as a part of the activities of the Science Advisory Board, a public advisory group providing extramural scientific information and advice to the Administrator and other officials of the Environmental Protection Agency. The Board is structured to provide a balanced, expert assessment of scientific matters related to problems facing the Agency. This report has not been reviewed for approval by the Agency; hence, the comments of this report do not necessarily represent the views and policies of the Environmental Protection Agency or of other Federal agencies. Any mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ABSTRACT

The Secondary Data Use Subcommittee of the Science Advisory Board's Executive Committee reviewed the Agency's draft "Data Suitability Assessment of Major EPA Databases". This assessment examines and reports upon the extent to which individual EPA regulatory databases can be used for a range of uses other than the use for which the database was designed. The Suitability Assessment is being performed in several stages of which the first, qualitative review, has been completed for six databases.

The Subcommittee found that the Data Suitability Assessment is appropriate for evaluating the general suitability of databases for a range of secondary uses. There was also a consensus that additions to what is in the present draft would improve the usefulness of the data bases to secondary users. The subcommittee not only recommended additions to the assessment but also suggested documents and activities beyond the assessment that would help researchers and the public understand the appropriate secondary uses of specific regulatory databases.

Key Words: data use; database

U.S. ENVIRONMENTAL PROTECTION AGENCY
SCIENCE ADVISORY BOARD
SECONDARY DATA USE SUBCOMMITTEE

CHAIR

Dr. Morton Lippmann, Professor, Nelson Institute of Environmental Medicine, New York University School of Medicine, Tuxedo, NY

MEMBERS

Dr. Miguel F. Acevedo, Professor, Institute of Applied Sciences and Department Geography, University of North Texas, Denton, Texas

Dr. Philip K. Hopke, Dean of the Graduate School and Professor of Chemistry, Clarkson University, Potsdam, NY **(Not available for December 1998 meeting)**

Dr. John P. Maney, President, Environmental Measurements Assessment, S. Hamilton, MA

Dr. Paul J. Merges, Chief, Bureau of Pesticides and Radiation, Division of Solid and Hazardous Materials, NY Department of Environmental Conservation, Albany, NY

Dr. Maria T. Morandi, Assistant Professor, University of Texas, School of Public Health, Houston, TX

Dr. Edo D. Pellizzari, Vice-President, Analytical and Chemical Sciences, Research Triangle Institute, Research Triangle Park, NC

CONSULTANTS

Dr John C. Bailar III, Chair, Department of Health Studies, University of Chicago, Chicago, IL **(Not available for December 1998 meeting)**

Dr. Richard O. Gilbert, Staff Scientist, Battelle Washington Office, Washington DC

Dr. Manuel Gomez, Director of Scientific Affair, American Industrial Hygiene Association, Fairfax, VA

Dr. Kinley Larntz, Professor Emeritus, University of Minnesota, Shoreview, MN

Mr. Douglas Splitstone, Splitstone & Associates, Murrysville, PA

SCIENCE ADVISORY BOARD STAFF

Mrs. Anne Barton, Designated Federal Officer, Environmental Protection Agency, Science Advisory Board, Washington, DC

Ms. Priscilla Tillery-Gadson, Program Assistant, Environmental Protection Agency, Science Advisory Board, Washington, DC

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The Science Advisory Board (SAB) reviewed the draft Data Suitability Assessment dated November 11, 1998 prepared by the Center for Environmental Information and Statistics (CEIS). This is a part of the ongoing effort by the SAB to help EPA meet the challenges it faces from release of Agency databases for secondary use. The review was done by the Secondary Data Use Subcommittee ("SDUS" or "the Subcommittee) of the SAB's Executive Committee at its meeting on December 15, 1998.

The EPA's Center for Environmental Information and Statistics (CEIS) is in the process of assessing major EPA regulatory databases for their potential use in secondary data analyses, (i.e., for uses other than those for which they were originally collected). There are several stages to this assessment of which the first, a *descriptive profile* has been completed for six EPA regulatory databases.

The Subcommittee found that the Data Suitability Assessment was carefully and thoughtfully assembled, and that it was an excellent first step in a process that should lead to a more widespread and productive utilization of historic EPA datasets. The Subcommittee members agreed that what had been included in the draft CEIS document is appropriate for evaluating the general suitability of databases for a range of secondary uses. There was also a consensus that additions to the present draft would improve the usefulness of the data bases to secondary users.

The Subcommittee's responses to the specific questions in the Charge are summarized below:

a) Are the descriptors included appropriate for evaluating the general suitability of the database for a range of secondary uses?

   The subcommittee agreed that the descriptors used in the draft are appropriate for evaluating the general suitability of the databases for a range of secondary uses.

b) Did CEIS handle the evaluation of these descriptors appropriately?

   Generally, the Subcommittee thought the handling of the descriptors was appropriate. The subcommittee made suggestions for improving the handling by bringing in more quantitative information; using terminology consistent with the Agency's guidance for (primary) Quality Assurance; drawing upon the results of the primary quality assurance process; providing more information on process: and otherwise enriching the descriptions. Much of this additional information may be added by the *statistical profile*, which will be the second step in the CEIS assessment.

c) What, if any, areas were missed in the suitability assessments that would make them more meaningful for scientific purposes?

The subcommittee suggested that the meaningfulness of the assessments would be enhanced by the addition of a glossary; qualitative indicators of data precision and bias (with caveats); the addition of information that may be available from additional sources; and indications of unreliable uses.

d) What specific questions would assist in further characterizing a database or set of databases for secondary uses?

The Subcommittee suggested questions concerning the primary purpose of the database (to what extent is it achieved, answers to questions from the Quality Assurance planning process); the design of the data collection effort; proportions of nondetects; how others have used the data base; representativeness of the data for particular secondary uses; whether uncertainty estimates are available; and other examples of questions that might assist secondary users to understand the strengths and limitations of the database.

e) What other advice does the Subcommittee have about ways to improve the usefulness of the descriptive review?

The Subcommittee recommended these activities in addition to the further development of the Suitability Assessment document:

(1) A Literature Review to assess the extent that the EPA databases have already been used for secondary purposes.

(2) A Guideline on Secondary Data Analyses in a format similar to that used in the preparation of its Exposure Assessment and Risk Assessment Guidelines but aimed, in this case, at potential secondary users of the databases

(3) A mechanism to provide feedback from users of the various databases.

The Subcommittee believes that these suggestions and changes will make the Data Suitability Assessment a more useful and transparent description of EPA databases and will improve their usefulness for secondary analyses of environmental quality conditions and trends.

2. BACKGROUND AND CHARGE FOR CEIS DRAFT ASSESSMENT

The SAB reviewed the draft Data Suitability Assessment dated November 11, 1998 prepared by the CEIS. This is a part of the ongoing effort by the SAB to help EPA meet the challenges it faces from release of Agency databases for secondary use. The review was done by the Secondary Data Use Subcommittee of the SAB at its meeting on December 15, 1998.

The EPA's CEIS is in the process of assessing major EPA regulatory databases for their potential use in secondary data analyses, (i.e., for uses other than those for which they were originally collected). There are several stages to this assessment:

a) A *descriptive profile* of each database derived from a questionnaire completed by the Program Office that maintains the database.

b) A *statistical profile* providing a quantitative characterization of key aspects of each database being reviewed.

c) A review of specific data applications.

The draft Data Suitability Assessment reviewed by the Secondary Data Use Subcommittee is based upon the *descriptive profile* of six EPA databases. It contains a "Major Findings Document" for each of the databases reviewed. These Major Finding Documents contain sections that address the spatial and temporal attributes of the databases and look at the ability to integrate the databases temporally and spatially. The Assessment is intended to help potential users evaluate the databases' general suitability for a range of secondary uses.

CEIS has completed a draft descriptive review of a subset of six of the Agency's regulatory databases and asked for SAB advice on the adequacy of this assessment as it is presently designed. Specifically:

a) Are the descriptors included appropriate for evaluating the general suitability of the database for a range of secondary uses?

b) Did CEIS handle the evaluation of these descriptors appropriately?

c) What, if any, areas were missed in the suitability assessments that would make them more meaningful for scientific purposes?

d) What specific questions would assist in further characterizing a database or set of databases for secondary uses?

e) What other advice does the Subcommittee have about ways to improve the usefulness of the descriptive review?

3. OVERVIEW OF CEIS DRAFT ASSESSMENT

The Secondary Data Use Subcommittee congratulates the Agency on undertaking this suitability assessment. It will be of great value to the Agency, to researchers outside the Agency, and to the general public. This is an extremely important aspect of making information available.

The Subcommittee found that the Suitability Assessment was carefully and thoughtfully assembled, and that it was an excellent first step in a process that should lead to a more widespread and productive utilization of historic EPA datasets collected largely for regulatory purposes. The Subcommittee agreed that what had been included in the draft CEIS document is appropriate for evaluating the general suitability of databases for a range of secondary uses. There was also a consensus that additions to what is in the present draft would improve the usefulness of the data bases to secondary users. Several kinds of additions are suggested in the following discussion of specific comments:

a) Additions to the present qualitative review of the databases (e.g., a glossary).

b) More quantitative additions that would be appropriate during the second, quantitative stage of review which CEIS plans for the future (e.g., quantitative precision information).

c) Activities or documents in addition to the assessment of the data bases (e.g., provision for user feedback).

The Subcommittee did not attempt to prioritize these recommendations but focused on providing a range of ideas and recommendations. This approach seemed likely to be more helpful to the CEIS at this point. However, the Subcommittee recognizes that priorities might be useful to the Agency as it decides how to uses limited resources in the immediate future. The Subcommittee would be happy to consider additional discussions to address priorities, if that would be helpful, while, at the same time, recognizing that factors other than scientific considerations will legitimately affect the Agency 's final decisions on priorities"

4. SPECIFIC COMMENTS ON CEIS DRAFT

4.1 Are the descriptors included appropriate for evaluating the general suitability of the database for a range of secondary uses? What additional areas would make the assessments more meaningful for scientific purposes? [Charge questions 1 and 3]

The Agency has done a very good job of compiling the basic information necessary for evaluating the suitability for secondary use of six databases examined to date. The assessment is clearly written and provides supplementary sources of information and contacts that the potential users of these data might need.

The type of descriptors of the database are appropriate for evaluating such suitability. However, the descriptors are not sufficient for achieving the stated goal. Additional information is needed to allow the user to decide the appropriateness of a database for a specific secondary use. (In addition to the specific suggestions given here in section 4.1, there are other suggestions later in this report. These include the documentation links suggested in section 4.2.6 and additional questions in section 4.3.

4.1.1 Need for a Glossary

The purpose of the descriptive suitability assessment is to communicate essential background information regarding EPA databases to the potential user. The descriptive profile found in the Major Findings document is the first level assessment that conveys to the user the suitability of the database for analyses beyond its primary purpose. As such, the descriptive profile must communicate this assessment in a manner that minimizes the potential for multiple or vague interpretations.

An important consideration in how the descriptors are presented is the nature of the customers of the database, which in this case spans the range from the lay public to the highly trained scientist. The lay user may not be familiar with the terminology in the suitability assessment and the nuances of the concepts behind these terms. The scientists at the other end of the spectrum may have a relatively narrow perspective of understanding from their particular discipline.

In order to improve the level of clarity for all potential customers, it is recommended that the Agency develop and provide a Glossary that defines technical terms as used in the Data Suitability Assessment. A few candidate terms are: accuracy, bias, precision, blunders and errors, internal consistency, representativeness, spatial and temporal characteristics, linkage, integration, and comprehensiveness (or completeness). The Glossary should not be limited to these, however.

In some cases, these definitions might be consistent with those already available and used by other Agency offices/programs, and an appropriate reference to the relevant document should be provided in such cases. (See Section 4.2.2.1 for some suggestions.) In other cases, the term may

be applied to a range of parameters (for example, "precision and accuracy" apply to the full range of data gathering activities, from sample collection to data entry and verification) and the glossary should present an explanation of the broad meaning of the descriptor or term together with the more narrow definition applied in the assessment. There may be a need for providing two levels of definitions, the first directed at the more technically trained customer, and the second to the lay public.

In addition to the explicitly stated data quality terms, additional terms may be included in the Glossary that serve the purpose of raising the users consciousness when considering the databases for a given secondary purpose.

4.1.2 Qualitative indicators of data precision and bias

Even though the Statistical Profiles (to be developed by CEIS) will provide information on data quality, it is important for the user to have a general sense of the level of precision and bias of the data in the Suitability Assessment since this could be an important consideration in his/her decision to use a particular database. As the Descriptive Profiles state, the reliability of the data varies both within and between databases, but the user does not have a sense of "how bad (or good)" the data can be. It is suggested that the Agency consider the development of "qualitative indicators" of data quality (e.g., high, moderate, low) for each database. Although these indicators are necessarily somewhat subjective and cannot fully describe the quality for any specific secondary use, they may be helpful to a potential secondary user by providing some general sense of the quality.

Criteria for this classification should include the extent of Quality Assurance/Quality Control (QA/QC) of the data prior to its incorporation in the database. For example, criteria pollutant concentration data in the Aerometric Information Retrieval System (AIRS) undergoes QA/QC following EPA guidelines in most cases, and could be considered in the high category (with exceptions noted for localities where those procedures are not followed). On the other hand, the extent of QA/QC for data in the Safe Drinking Water Information System (SDWIS) is highly variable and more difficult to ascertain, probably in the low to moderate category (again, with exceptions noted where necessary for particular localities). In the case of the Toxics Release Inventory (TRI) estimates, there are different levels of accuracy depending on the specific compound and the type of source, so the qualitative indicator could be assigned on the basis of how accurate the estimate is expected to be for a particular compound or class of compounds. There may be also varying accuracy depending on the reporting facility and others as indicated on page 17 of the TRI description.

These qualitative indicators for the precision and bias of a database should include a caveat that the acceptability of the database will have to be re-evaluated in terms of the proposed secondary data use since the reliability of data is dependent upon their use and will vary according to the secondary data use.

4.1.3 Other sources

Needed information for particular secondary uses may be available in other sources besides the specific database so that a limitation for a specific use could be addressed with information available elsewhere. For example, TRI does not contain information on utility emissions, or mobile sources. However, there are other databases (the Acid Rain Program's Emission Tracking System (ETS) in the case of utilities or state implementation plans in the case of mobile sources) that can provide at least some of this information. The Subcommittee recommends that the Agency provide links to other sources of reliable information that the user can access to supplement the data limitations of the various databases.

4.1.4 Unreliable uses

Each Major Findings Document describes the types of analysis (e.g., temporal, spatial) that can be done with the six databases. In the case of the Aerometric Information Retrieval System data, a descriptor of the selection criteria for sampling site location would be useful for deciding if a secondary analysis can be done. It would be useful also to add descriptors of unreliable uses. For example, the Safe Drinking Water Information System data have very limited utility for estimating concentrations of regulated chemicals in water at the household level. In the latter case also, data are reported as in compliance or as a concentration if the sample exceeded the regulatory benchmark. In these cases, there should be also information on the benchmark concentration (which could vary over time).

4.2 Did CEIS handle the evaluation of these descriptors appropriately? [Charge question 2]

4.2.1 Spatial and Temporal Analysis

CEIS has done a nice job of describing the available time and/or space location measures for the raw data. This information, where available, appears to be sufficient to initiate an analysis of spatial and/or temporal patterns among the data. The Data Suitability Assessment clearly identifies the possibility of performing or not performing these analyses along with the description of the relevant parameters.

However, several of the databases contain summarizations of the raw data (e.g., time averages, spatial trends) without providing the basis for their development. In order to determine the suitability for secondary use of these data summary statistics, one must have knowledge of the time and/or space lattice of individual data points used in their construction. Full and complete explanations of these data summarization processes, including the rationale for datum selection and algorithms for summarization, should be included to determine the suitability for secondary use of these summary statistics. Such information may be included in the Statistical Profile stage of the assessment.

Location information is the major link across databases. This is, therefore, a variable that should be highly reliable. At least with TRI, there are errors in the source locations. It is recommended that the Agency address such errors to the largest possible extent.

4.2.2 Accuracy of the Data

The following comments, recommendations and findings with regard to the accuracy and limitation aspects of the Suitability Assessment are based on the presentations made by Agency personnel and the documentation that was distributed to and reviewed by the Subcommittee prior to its meeting in December 1998. The Subcommittee recognizes that completion of the statistical profile portion of the Suitability Assessment may result in changes that could alter the Subcommittee's recommendations and findings.

4.2.2.1 Definitions

There are conflicting definitions for commonly used terms such as "accuracy" and numerous synonyms for the sampling and analytical errors that can affect data quality. To minimize confusion among the readers of the Major Finding documents, the Subcommittee recommends that the Agency's definition for "accuracy", "representativeness" and the three error types; mistakes, bias and imprecision be referenced or included in the Accuracy of Data sections and in the glossary. Some definitions are suggested below. All but the last of these are from the EPA Quality Assurance Guidance Document G-5, published in August, 1997.

a) accuracy - A measure of the closeness of an individual measurement or the average of a number of measurements to the true value. Accuracy includes a combination of random error (precision) and systematic error (bias) components that are due to sampling and analytical operations; the EPA recommends using the terms "*precision*" and "*bias*", rather than "*accuracy*" to convey the information usually associated with accuracy.

b) representativeness - A measure of the degree to which data accurately and precisely represent a characteristic of a population, a parameter variation at a sampling point, a process condition or an environmental condition.

c) bias - The systematic or persistent distortion of a measurement process, which causes errors in one direction (i.e., the expected sample measurement is different from the sample's true value).

d) precision - A measure of mutual agreement among individual measurements of the same property, usually under prescribed similar conditions expressed generally in terms of the standard deviation.

e) blunders and errors - are mistakes such as transcription errors, which occur on occasion and cause erroneous results (John Taylor, "Quality Assurance of Chemical Measurements")

The Questionnaire for databases uses the phrase "precision and accuracy". This use equates "accuracy" to "bias" which is in conflict with the Agency's QA/QC definitions. The Subcommittee suggests that CEIS harmonize its definitions with those detailed in the Agency's Quality Assurance documents. The use of harmonized definitions will allow the assessment to discriminate between different error sources such as measurement errors and transcriptions errors, which will facilitate evaluation of the database for secondary uses.

4.2.2.2 Description of evaluation process

In addition to addressing the above definitions, the Subcommittee recommends that the "Accuracy of Data" section describe data evaluation processes in more detail so the potential data user will be able to interpret the data evaluation process in light of a specific secondary-data use. The descriptions should provide sufficient detail, include the outcome of the evaluation process, and interpret the outcome in terms of accuracy. For example, in the respective Major Findings documents for individual databases;

a) The discussion in the “Accuracy of Data” section of CEIS’s Major Findings Document on the Toxics Release Inventory would be improved by a list of the Standard Industrial Code (SIC) for the audited industries, and a description of how facilities within a SIC were selected. Was the audit focused on large facilities? Small facilities? Problematic facilities? A cross-section of facilities?

b) Although 5% of all Aerometric Information Retrieval System - Air Quality Subsystem data have been audited, the type and depth of the audit is undefined. The term "gross errors", although defined in the questionnaire, is not defined in the Major Findings document. There were some detailed precision and accuracy data included in the questionnaire for this data base which should be included in its Major Findings document.

c) The Major Finding document for the Acid Rain Program’s Emissions Tracking System Database indicates a gross error rate of approximately 1% and an availability (uptime versus downtime) of 96% for monitoring systems. Some discussion of the bias or precision of the actual SO_2, NO_x and CO_2 measurements would be useful.

d) The Major Findings Document for the Safe Drinking Water Information System indicates a 12% error rate for significant non-compliance, yet it does not indicate whether this non-compliance rate is stratified by contaminant or by size of supplier. If the accuracy of data varied according to contaminant, the size of the supplier or

some other variable, this information may be useful in determining the applicability of the database for secondary data use. Neither the Major Findings document nor the questionnaire offered any bias or precision information.

e) A stated goal for the Permit Compliance System database is to be 95% confident that the actual pipe position is within 25 meters of the reported location, however there is no indication as to whether this goal is met. The Major Findings document indicates that "QA/QC procedures are in conformance with EPA requirements". It would be helpful to potential secondary users to indicate whether these are procedures for sampling, analysis, or data reporting. This statement indicates that QA/QC procedures are in place, but does not indicate whether they are implemented appropriately nor does it describe the results of their implementation. It is unclear as to whether the procedures succeed in documenting the quality of the data that reside in this database. This additional information would be helpful.

f) The Resource Conservation and Recovery Act Information System (RCRIS) Major Findings document indicates that the EPA Regional inspectors check 10% of the state-implemented inspections. However, there are no indications of what the Regional inspectors found and the impact of these findings on the quality of the RCRIS data. The questionnaire indicates that a GAO audit of RCRIS was "generally unfavorable", some indications as to how the GAO findings may impact data quality would be useful.

Greater detail regarding the evaluation and audit process as well as the quantitative accuracy and precision information would greatly facilitate the assessment of the above databases for secondary data uses. The Subcommittee recognizes that the CEIS Statistical Profile (now in preparation) may fill many of these gaps.

4.2.2.3 Primary data quality

The Subcommittee recognizes that data quality issues are a function of the objectives of a study and are likely to vary from primary to secondary data use. In particular, this is the case for the representativeness of samples and the associated data. Recognizing this study-dependent aspect of representativeness, the Subcommittee recommends that the Agency discuss within the assessment the key issues that affect representativeness. For example, for a contaminant-measurement database, how sampling locations were identified (i.e, the sampling design - e.g., simple random or biased sampling designs), how samples and subsamples were collected (i.e., sample support issues such as sample size, sample mass/volume and orientation). This information will allow the secondary-data user to evaluate the representativeness of the data in terms of the specific secondary data use being considered.

Regarding the accuracy of Agency databases, it should be noted that environmental laws were enacted and regulations authored with the idea that the Agency would typically rely upon

the States, the regulated community and others to assist in the implementation of Congress's intent. All of the databases subjected to the data suitability assessment rely heavily upon input from the States and the regulated community. A recent EPA Scientific Advisory Board review[1] uncovered that "more than 75% of states lack approved quality management plans for all or significant numbers of their environmental programs …… States lacking a Quality System for environmental programs are unlikely to document the quality of data..... exposing itself, the reliability of its decisions, and its credibility, to criticisms due to reliance upon data of unknown quality. The same is true for those Agency programs, which depend upon those data."

Likewise, anecdotal information uncovered during the Quality System review indicated that data generated by other organizations, such as the regulated community, also lack an approved Quality System. The Subcommittee finds that the usefulness of databases generated by the Agency and other parties will be negatively impacted until the Agency implements its Quality System across the Agency and establishes a quality assurance mechanism for oversight of those who have data collection responsibilities for Agency's environmental programs.

4.2.3 Limitations of the data

The Subcommittee recommends that the limitations section discuss or reference other limitations that are detailed in other sections of the Major Findings document for a given database.

Limitations will be relative to the objectives of any given secondary data use. Limitations for the original study may or may not be limitations for the secondary data use. Conversely, non-issues for the original data use may prove to be limitations to a secondary data use. While it is not possible to anticipate all secondary data uses, a thorough documentation of known limitations and a complete description of the database, its purpose and contents should assist those anticipating secondary use.

To be more thorough, the limitation section should reference those limitations discussed in other sections of the Major Findings document. Although the content of the accuracy and spatial and temporal analyses sections do not need to be repeated they should be referenced. A complete inventory of limitations in one section will uncover pertinent issues and direct the reader to pertinent issues even during a cursory review.

4.2.4 Comprehensiveness

The Subcommittee is supportive of the Agency's intent to define and report the comprehensiveness of each database profiled. The Subcommittee understands that the term "comprehensiveness" is being used to signify the scope of coverage of each database, as well as

[1] "The Science Advisory Board Review of the Implementation of the Agency-Wide Quality Management Program", by the Environmental Engineering Committee of the Science Advisory Board, US EPA, Washington DC. EPA-SAB-EEC-LTR-99-002, February 22, 1999.

the extent to which it is achieved in practice. For example, the comprehensiveness of the TRI database is qualitatively described in the current profile by making clear that the toxic emissions covered reflect only certain industry sectors, that non-point source emissions are not included, etc. A clearer definition of the use of the term in these profiles would be useful, possibly with a clarifying example.

4.2.5 Links

The Subcommittee supports the emphasis on describing the links in each database that permit it to be connected to other databases for combined analyses. Links are understood by the Subcommittee to refer to data elements that are common to two or more databases. Five types of links were considered critical for description. These links were those related to: a) geographic location (lat/long, etc.); b) time (hour, date, etc.); c) media; d) CAS number; and e) unique identifying number for establishments

The Subcommittee strongly supports the Agency's efforts to assign unique identifying numbers to establishments that are used by all relevant databases in the Agency. To the extent that all six databases contain compound-specific information, the CAS number for each compound can also be used as a link among the databases.

4.2.6 Documentation

The Subcommittee is very supportive of steps that would encourage investigators and analysts to attempt to use the databases for appropriate secondary purposes. For this reason, it strongly recommends that the Agency take steps to make it as easy as possible for potential users to access both the data and the documentation that describes each one. The Subcommittee understands that, at present, the databases may be difficult to access even for sophisticated potential users. This obstacle should be addressed. The database profiles should be linked, as much as possible, to Web pages containing detailed documentation for each database, as well as knowledgeable contact persons. A useful approach may be for the agency to design the profiles as screening tools with a cascading degree of information for increasingly sophisticated users.

4.3 What areas or specific questions would make the assessments more meaningful for scientific purposes? [Charge question 4]

As discussed in section 4.1, the set of nine questions used in the Assessment are appropriate and appear to elicit good information about the various databases. However, in addition to the issues listed in section 4.1, there are additional questions that should be considered for obtaining more detailed information. In particular:

a) A question should be asked to determine if the primary purpose of the database is being achieved to the needed extent. Are the primary purposes of the databases clearly defined, and to what extent is the database capable of fulfilling them? What

characteristics of the database have proved to be obstacles to achieving this primary purpose? If necessary, follow-up questions should be asked to determine what problems are preventing the full achievement. A summary of whatever difficulties have been identified would help inform the evaluation of the usefulness of each database for secondary purposes. This question of "primary purpose" would complement and expand upon the question "What Does the Database Cover?"

b) Additional questions should be asked to obtain more information on the design of the data collection effort (where, when, and how environmental samples were collected, handled, and measured), as well as on the different data elements obtained by the design. The user of the database for secondary purposes must know about the design because it will have a significant impact on whether the data are suitable for the envisioned secondary use.

c) A question should be added that elicits information about the proportion of the data in the database that are non-detects, i.e., measurements that are less than the detection limit and how less-than values are represented in the database. The number of secondary uses of a database may be drastically reduced if most data are less-than values or blanks or zeros.

d) The following question should be added: "How have others successfully or unsuccessfully used the database for secondary uses?" As described in Section 5, below, the Subcommittee feels that the Agency, the profiles, and the eventual users of the data would benefit greatly by the completion of a literature search regarding secondary uses that have been attempted, and the difficulties they have encountered. The results of this literature search, possibly with some commentary, would be a useful addition to the profiles. What is the range of secondary uses that have been attempted for each database? To what extent have these attempts been successful? Why or why not?

e) There was considerable discussion among the Subcommittee members about the need for representative data for secondary data uses. The concept of "representative data" is somewhat complex; it may be difficult to devise a suitable question unless the particular secondary data use is well defined. Nevertheless, we encourage EPA to attempt to develop a question or series of questions that will provide information to potential users about the appropriateness of the data for different purposes. The EPA might address this problem by developing several examples where, for a given secondary data use, an appropriate question(s) about representativeness is provided.

f) A question that should be considered for inclusion is whether the database reports an uncertainty or variance value for each measurement, and more generally,

whether there are measures of variance or uncertainty in the database for reported data summaries (e.g., averages). Also, some databases contain estimates of parameters rather than actual measurements of those parameters. The estimates may be computed in various ways. Do any of the databases provide uncertainty bounds on those computed estimates?

g) The EPA requires that the Data Quality Objectives (DQO) process be used to plan the collection of data to ensure that the data obtained are of the quantity, quality and type required for the primary purpose of the database. This planning process is intended to produce a suitable sampling design (where, when, what). One or more questions that elicit information about the extent to which the 7 steps of the DQO process were used may be helpful in establishing the quality of the data. These questions would not focus on the quality of the measurements, but rather on the quality of the sampling design.

h) There are many specific questions that could be asked. The following list is offered not as a "must use" list of questions, but to stimulate thought about which, if any, of these questions are not being adequately addressed with the current list of nine questions.

(1) Who reports the data?
(2) Who funds the data collection effort?
(3) Are some data useless?
(4) Are spurious data reported?
(5) Is the ZIP code reported accurate and meaningful?
(6) Is data considered to be interim or final?
(7) Will a final report be issued?
(8) Were samples collected at suspected hot spots?
(9) Were samples collected using a probability-based sampling plan such as simple random sampling?
(10) Were samples collected at locations on a square, rectangular, or triangular grid?

In addition to questions asked to elicit information about databases from EPA program offices, we suggest that EPA consider developing a set of generic questions that potential data users could answer to help them assess their own data needs and whether the database is suitable for that purpose. We envision that the potential data user who accesses the EPA web site would be encouraged to spend 15 minutes to try and answer them. Moreover, EPA should devise a method, if needed, for these users to contact EPA personnel via the web or other means to help the user answer the questions. The key concern is that users will need help in thinking about their data needs and whether the database is appropriate for their particular need. This concept is closely tied to the idea, discussed elsewhere in this report, of EPA developing guidelines for

secondary data users on how to evaluate whether a database contains information of the necessary form and quality for the intended secondary data use.

4.4 Other considerations and ideas

4.4.1 Literature Review

The Subcommittee recommends that the CEIS perform a literature review to assess the extent that the EPA databases have already been used for secondary purposes. The results of this study should identify reports in which researchers have accessed the EPA databases; how any links to other databases were made; how EPA's data were used in a secondary mode; and the results of this secondary use. CEIS should discuss these "live" case studies with the authors of these reports and develop a "lessons learned" report from these studies. CEIS can then factor these results into its assessment of secondary use of data as appropriate. As the information program evolves, this literature review should be updated. The results should provide useful information to improve the secondary use assessment. By bringing the review results to SDUS in some future meetings, CEIS could also help the Subcommittee accomplish its charge.

4.4.2 Guidelines for Secondary Data Analysis

CEIS's draft "Data Suitability Assessment of Major EPA Databases" is an excellent first step in providing guidance to the general public on the potential uses of the EPA databases. In addition, we recommend the development of a Guideline on Secondary Data Analyses in a format similar to that used in the preparation of its Exposure Assessment and Risk Assessment Guidelines but aimed, in this case, at potential secondary users of the databases. This assessment would provide step-by-step technical and specific guidance to the user on how to approach using the database and would contain more details on statistical assessments, data quality, measurement analysis, as well as technical and scientific references. Such guidance should specify the steps of the Data Quality Objectives Process (EPA QA/G-4) applicable to conducting post hoc data analyses and take full advantage of the discussion of strengths and weaknesses of various statistical methods found in the Guidance for Data Quality Assessment (EPA QA/G-9).

Such an assessment would facilitate and encourage productive application of EPA's data resources by both Agency and outside investigators and help channel such analyses into analytical frameworks and formats that will be more uniform and useful for integration into future Agency programs, analyses, and decision assessments. Also, it would be helpful to users if the step-by-step planning process described in the guidelines document was illustrated using 3 or 4 different types of secondary data uses. An example of a misuse of the database(s) would also be helpful. A secondary data user should be guided on the process he/she should use to access, understand, extract and use the specific data suitable for the particular data use in mind.

In addition, these guidelines could be incorporated into an online step-by-step "navigator" linked with web sites related to the databases, as well with other sites. These linkages will enrich the

guidance provided to the user. CEIS is already planning the development of such a project. We endorse pursuing this possibility.

4.4.3 User Feedback

The Subcommittee recommends that a mechanism be developed to provide feedback from users of the various databases. The primary purpose of getting such feedback is to provide information about using the databases for other future users. An additional purpose is provide information to EPA for improving the usability of the databases for secondary data analyses. A third purpose is to inform the interested public about the range of uses of and potential substantive results from the databases.

There will likely be difficulties in data handling as well as issues concerning the data collection design and the data elements themselves that could be usefully shared. It would also be a place for users to post questions that more experienced users could answer. The posted answers would then be available for all database users.

It is recommended that EPA review the comments for suggestions on improving the documentation for the database. Actual users will likely test the database more thoroughly and more completely than any planned agency examination.

Finally, users may post the results of their analyses to share with other interested parties. In some sense, the other interested parties could serve as peer reviewers of the posted results. It is anticipated that active discussions could arise and that other secondary analyses be stimulated from such postings.

The Subcommittee recommends that the method for user feedback be the filing of a user's report online or by E-mail. A box soliciting user comments could be put on the web page for the Major Findings document for each database. The solicitation box would encourage users to provide comments for all three purposes given above. The Subcommittee recommends that all signed comments be posted on the web site. Initially, all postings could be placed in a single assessment , and, perhaps, all new postings should be in a single assessment. When this single assessment becomes large enough, it will be necessary to archive the postings. Given the three purposes for user feedback, the achieved postings could be divided into: a) hints for other users; b) suggestions for improvements; and c) secondary data analysis results.

The Subcommittee believes that these suggestions and changes will make the Data Suitability Assessment a more useful and transparent description of EPA databases and will improve their usefulness for secondary analyses of environmental quality conditions and trends. It looks forward to future opportunities to help EPA in its effort to improve the analyses of the influence of environmental quality on human health, welfare, and ecological systems.

APPENDIX A - GLOSSARY

Accuracy – A measure of the closeness of an individual measurement or the average of a number of measurements to the true value. Accuracy includes a combination of random error (precision) and systematic error (bias) components that are due to sampling and analytical operations; the EPA recommends using the terms "*precision*" and "*bias*", rather than "*accuracy*" to convey the information usually associated with accuracy. (From EPA QA/G-5)

AIRS, AIRS-AFS – Aerometric Information Retrieval System.; Facility Subsystem. A database which contains air pollutant compliance, permit, and emissions data for stationary sources regulated by the EPA, and State and local agencies.

AQS – Air Quality Subsystem. A subsystem of AIRS containing information about certain types of airborne pollutants in the United States and various World Health Organization member countries.

Bias – The systematic or persistent distortion of a measurement process, which causes errors in one direction (i.e., the expected sample measurement is different from the sample's true value). (From EPA QA/G-5)

Blunders and Errors – Mistakes such as transcription errors, which occur on occasion and cause erroneous results. (From John Taylor, "Quality Assurance of Chemical Measurements")

CAS number – An identification number assigned to a particular chemical substance by the Chemical Abstracts Service (CAS) Registry.

CEIS – Center for Environmental Information and Statistics. This is the EPA organization which has taken on the role of conducting Data Suitability Assessments of EPA data bases.

Data Suitability Assessment – An assessment of the degree to which major EPA databases can meet the varying demands of a wide range of information users. There are several components to these assessments, including a Descriptive Profile and a Statistical Profile.

Descriptive Profile – A description of a database derived from a questionnaire completed by the Program Office that maintains the database.

DQO – Data Quality Objectives – The qualitative and quantitative statements derived from the DQO Process that clarify study's technical and quality objectives, define the appropriate type of data, and specify tolerable levels of potential decision errors that will be used as the basis for establishing the quality and quantity of data needed to support decisions. (From EPA QA/G-5)

DQO Process – Data Quality Objectives Process – A systematic strategic planning tool based on the scientific method that identifies and defines the type, quality, and quantity of data needed to satisfy a specified use. DAOs are the qualitative and quantitative outputs from the DAO process. (From EPA QA/G-5)

EPA QA/G -5 – EPA's Quality Assurance Guidance Document which provides guidance on developing quality assurance plans. See **QA/G-**

ETS – Emission Tracking System. A database containing data collected under the auspices of the Acid Rain Program and the Clean Air Act.

Major Findings Document – A document prepared for each database included in the CEIS suitability assessment. Major findings documents will include results from descriptive profiles, statistical profiles, and other components of the assessment. The Major Findings Documents in CEIS's November, 1998 draft includes results of the descriptive profiles alone.

PCS – Permit Compliance System – A database of information on water discharge permits, designed to support the National Pollutant Discharge Elimination system.

Precision – A measure of mutual agreement among individual measurements of the same property, usually under prescribed similar conditions expressed generally in terms of the standard deviation.

QA/G - A series of EPA Quality Assurance Guidance Documents. QA/G-5 is the one of this series which provides guidance on developing quality assurance plans.

QA/QC – Quality Assurance and Quality Control

QA – Quality Assurance – An integrated system of management activities involving planning, implementation, assessment, reporting, and quality improvement to ensure that a process, item, or service is of the type and quality needed and expected by the client. (From EPA QA/G-5)

QC – Quality Control – The overall system of technical activities that measures the attributes and performance of a process, item, or service against defined standards to verify that they meet the stated requirements. The system of activities and checks used to ensure that measurement systems are maintained within prescribed limits, providing protection against "out of control" conditions and ensuring the results are of acceptable quality. (From EPA QA/G-5)

Quality System – A structured and documented management system describing the policies, objectives, principles, organizational authority, responsibilities, accountability, and implementation plan of an organization for ensuring quality in its work processes, products (items), and services. The quality system provides the framework for planning, implementing, and assessing work

performed by the organization and for carrying out required quality assurance (QA) and quality control (QC). (From EPA QA/G-5)

RCRIS – Resource Conservation and Recovery Act Information System. A database containing information on the identification, location, permitting status, closure/post-closure status, compliance, and enforcement issues for hazardous waste handlers regulated under the Resource Conservation and Recovery Act.

Representativeness – A measure of the degree to which data accurately and precisely represent a characteristic of a population, a parameter variation at a sampling point, a process condition, or an environmental condition.. (From EPA QA/G-5)

SDWIS – Safe Drinking Water Information System. A database containing information on drinking water contamination levels as required by the Safe Drinking Water Act.

SIC – Standard Industrial Classification. A code used to describe the type of work performed by a business establishment.

Statistical Profile – A quantitative characterization of key aspects of the database.

TRI – Toxics Release Inventory. A database containing information on releases of specific toxic chemicals, as required by the Emergency Planning and Community Right to Know Act.

DISTRIBUTION LIST

Administrator
Deputy Administrator
Assistant Administrators
EPA Regional Administrators
Director, Office of Science Policy, ORD
EPA Laboratory Directors
EPA Headquarters Library
EPA Regional Libraries
EPA Laboratory Libraries
Library of Congress
National Technical Information Service
Congressional Research Service