EXPOSURE DRAFT

CRITERIA FOR EVALUATING THE INTEGRITY OF A SET OF DATA

May 1, 2018

Comments are requested by July 9, 2018

Prepared by the AICPA Assurance Services Executive Committee's
ASEC Trust Information Integrity Task Force

1 2 3 4 5 6 7 8 9 0 BRAAS 1 9 8

Page

Explanatory Memorandum

Exposure Draft

Explanatory Memorandum

Introduction

We live in an era of data-based decision making, with data being provided through increasingly sophisticated processes, such as data mining. Data now plays a major role in decision making throughout entities in most industries, and the capability of performing data analytics is seen as a competitive advantage by many leading industry entities. Given this dependency, the quality of decisions made is directly affected by the quality of the data used; if the data supporting a decision is of poor quality, that decision may be erroneous or insufficiently informed. However, even the use of high quality data cannot prevent a poor decision.

A key aspect of data quality is its integrity. The AICPA has identified a need for criteria for evaluating the integrity of a set of data used in decision making as part of the profession's obligations to serve the public interest. The criteria can be used to assist senior management, boards of directors, internal auditors, and other stakeholders in assessing a set of data used in decision making. The criteria can also be used by a CPA in attestation and consulting engagements on the integrity of a set of data.

To reduce the risk that a set of data is misused or misinterpreted, the data needs to be accompanied by a description of the set of data, or a description, otherwise, needs to be available to the users. This allows a user to understand the nature of data within the set, its intended use, and information about what determines membership in the set. Accordingly, this information also enables the user to understand the limitations of the set of data.

For a set of data to have integrity, it needs to exhibit the following two attributes:

1. A description of the set of data presented in accordance with the criteria in this document is made available to the users.

2. The set of data is consistent with the description.

This document presents criteria for use when preparing a description of a set of data. In addition to the criteria, accompanying implementation guidance presents factors to consider when making judgments about the nature and extent of disclosures required by each criterion. The implementation guidance does not address all possible situations; therefore, users should carefully consider the facts and circumstances of the entity and its environment in actual situations when applying the criteria.

In establishing and developing these criteria, the Assurance Services Executive Committee (ASEC) follows due process procedures, including exposure of criteria for public comment. BL section 360R, *Implementing Resolutions Under Section 3.6 Committees*[1], designates ASEC as a senior technical committee with authority to make public statements and publish criteria without clearance from the AICPA council or the board of directors. Accordingly, ASEC will conclude whether the criteria are suitable criteria for preparing, and evaluating the presentation of, the description of a set of data.

---

[1] All BL sections can be found in AICPA *Professional Standards*.

The criteria proposed in this document are for use when preparing, or evaluating, a set of data, including when the practitioner is engaged to provide attestation or other nonattest or advisory services to a client. For instance, a practitioner may use such criteria when engaged to assist management with the development of its description of a set of data for use in internal reporting. The criteria can also be used to analyze the sufficiency of the description of a set of data prior to using that data.

Guide for Respondents

ASEC is seeking comments specifically on the nature and extent of information and disclosures contained in the proposed criteria. Specifically, respondents are asked to address the following questions:

1. Are any of the criteria or implementation guidance unnecessary or otherwise not relevant? Please provide a list.

2. Are there any missing criteria or implementation guidance? Please provide a list.

3. Do you have any concerns about the measurability of any of the criteria or implementation guidance? Please provide a list.

Comments are most helpful when they refer to specific paragraphs or criterion numbers, include the reasons for the comments, and, when appropriate, make specific suggestions for any proposed changes to wording. When a respondent agrees with proposals in the exposure draft, it will be helpful for the ASEC Trust Information Integrity Task Force to be made aware of this view, as well.

Written comments on the exposure draft should be sent to Di Krupica at di.krupica@aicpa-cima.com and received by July 9, 2018.

Comment Period

The comment period for this exposure draft ends July 9, 2018.

Assurance Services Executive Committee

(2017–2018)

Robert Dohrer, *Chair*
Bradley Ames
Christine M. Anderson
Nancy Bumgarner
Jim Burton
Mary Grace Davenport
Chris Halterman
Elaine Howle
Jennifer Haskell
Bryan Martin
Brad Muniz
Joanna Purtell
Miklos Vasarhelyi

ASEC Trust Information Integrity Task Force

Chris Halterman, *Chair*
Dennis Bell
Efrim Boritz
Charles Curran
Sheri Fedokovitz
Peter F. Heuzey
Kevin Knight
Christopher W. Kradjan
Dave Palmer
Tom Patterson
Robert Ramsey
Rod Smith
Jerry Trites

AICPA Staff

Amy Pawlicki
Vice President
Assurance and Advisory Innovation

Erin Mackler
Director
Assurance and Advisory Innovation, SOC Services

Di Krupica
Lead Manager
Product Management and Development
Public Accounting

Proposed Criteria for Evaluating the Integrity of a Set of Data

Introduction

1. *Data* is typically defined as facts and statistics collected about the characteristics or attributes of events and instances for reference or analysis or when used as a basis for further calculation, reasoning, discussion, and informing data-based decision making. A *set of data* is a defined collection of data regarding events that share common characteristics or relationships.

2. Collecting and transforming raw data into a set that is useful for decision making usually involves collecting a sufficient amount of relevant data, recording it, and then aggregating, stratifying, or grouping it for subsequent use. For example, information on the daily high and low temperatures for a community is determined by processing data from individual thermometers. Air temperature is continuously measured by a thermometer (data), but it may only be collected, read, and recorded once per minute, creating a set of data of the recorded temperature readings of that particular thermometer for that moment or a range of time based observation points. The data can then be processed with data from other thermometers, which vary in terms of location in the community or type of thermometer, to determine the daily high and low temperatures.

3. Data and sets of data have varying degrees of structure. They may be highly structured (for example, phone numbers), partly structured (for example, email and object-oriented databases), or unstructured (for example, a series of characters in a sentence or pixels used to create a television image). However, even highly structured data may have unstructured elements (for example, the description field in a journal entry), and data that appears to be unstructured may have structured elements (for example, the data stored in a computer file appears to be random characters unless the user knows how the data was formatted and stored).

4. The importance of the integrity of a set of data (for example, the data within the set is valid, complete, accurate, and current) will vary depending on the intended use. For example, airline pilots have a greater need for accurate data regarding wind speed at an airport than does a person deciding whether to wear a jacket. The degree of completeness and accuracy required in preparing the data is a matter of judgment (similar to the concept of materiality in a set of financial statements) depending on the intended use of the data.

5. To reduce the risk that a set of data is misused or misinterpreted, the data needs to be accompanied by a description of the set of data or a description is otherwise needed to be available to the users. This allows a user to understand the nature of data within the set, its intended use, and information about what determines membership in the set. Accordingly, this information also enables the user to understand the limitations of the set of data.

6. The AICPA Assurance Services Executive Committee (ASEC), through its Trust Information Integrity Task Force, has developed a set of benchmarks, known as *criteria*, to be used when preparing a description to help users understand data that an entity uses internally or provides to other third-party users. The criteria can be used to assist senior management, boards of directors, internal auditors, and other stakeholders in assessing the quality of data used in decision making. It may also be used by a CPA in an attestation engagement on the integrity of the data in accordance with the AICPA *Statements on Standards for Attestation Engagements* or in providing consulting services.

7. For a set of data to have integrity, it needs to exhibit the following two attributes:

a. A description of the set of data presented in accordance with the criteria in this document is made available to the users.

b. The set of data is consistent with the description.

8. Because management is ultimately responsible for the description of the set of data, it is management's responsibility to develop and present the description.

9. Management uses the criteria when preparing the description of a set of data; the CPA uses the criteria when evaluating whether the description is presented in accordance with the criteria and the set of data is presented in accordance with the description.

10. This document presents the criteria for use when preparing a description of a set of data and implementation guidance that represents important characteristics of the criteria.

11. In establishing and developing these criteria, ASEC has followed due process procedures, including exposure of criteria for public comment. BL section 360R, *Implementing Resolutions Under Section 3.6 Committees*, designates ASEC as a senior technical committee with the authority to make public statements and publish criteria without clearance from the AICPA council or the board of directors. Accordingly, ASEC will conclude whether these criteria are suitable criteria for preparing, and evaluating the presentation of the description of a set of data.

Professional Standards That Apply to Data Integrity Attestation Engagements

12. A data integrity attestation engagement is performed in accordance with AT-C section 105, *Concepts Common to All Attestation Engagements*[2], and other applicable sections of the attestation standards, depending on the type of attestation service being performed (for example, AT-C section 205, *Examination Engagements*). For example, a social networking website for professionals within an industry may wish to provide an anonymized set of data regarding network members in order to obtain financing. In such a situation, an examination on the integrity of the set of data may be useful in obtaining more favorable terms. Under the attestation standards, the CPA performing an attest engagement is known as a *practitioner*.

13. According to paragraph .A42 of AT-C section 105, the attributes of suitable criteria are as follows:

a. *Relevance.* Criteria are relevant to the subject matter.

b. *Objectivity.* Criteria are free from bias.

[2] All AT-C sections can be found in AICPA *Professional Standards*.

c. *Measurability.* Criteria permit reasonably consistent measurements, qualitative or quantitative, of subject matter.

d. *Completeness.* Criteria are complete when subject matter prepared in accordance with them does not omit relevant factors that could reasonably be expected to affect decisions of the intended users made on the basis of that subject matter.

14. In addition to being *suitable*, paragraph .25*b* of AT-C section 105 indicates that the criteria used in an attestation engagement should be available to users. The publication of this criteria makes it available to users.

15. In addition to their use as suitable criteria, responsible parties or organizations may also find the criteria in this document useful in developing criteria particular to a specific subject matter that are relevant, objective, measurable, and complete.

16. The criteria in this document also may be used when the practitioner is engaged to provide other nonattest or advisory services to a client. For instance, a practitioner may use the criteria when engaged to assist management with the development of its description of a set of data for use in internal reporting. Consulting services are performed in accordance with the guidance in CS section 100, *Consulting Services: Definitions and Standards*[3].

17. If the practitioner assists management with the development of its description, threats to the practitioner's independence may exist. The "Nonattest Services" interpretation (ET sec. 1.295)[4] of the "Independence Rule" (ET sec. 1.200.001) provides special independence requirements for practitioners who provide nonattest services for an attest client. In addition, the "Conceptual Framework Approach" (ET sec. 1.210) discusses threats to independence not specifically detailed elsewhere.

Preparing a Description of a Set of Data

18. A description of a set of data is intended to provide users of the data with context that will enable them to understand the data and make appropriate decisions based on that data.

19. Included in paragraph 24 are the criteria for use when preparing a description of a set of data. For the description to be presented in accordance with the criteria, the context and disclosures addressed by the criteria should be included in the description. In addition, each of the relevant elements required by the criteria should be adequately described or disclosed (for example, sufficient to avoid the description from being misleading; for example, there are material omissions or misrepresentations). For some types of elements or sets of data, there may be specific, defined criteria that are relevant in the determination of the population, the nature of each element, the source of the data, units of measurement, accuracy, correctness or precision of measurement, uncertainty or confidence in the elements, or other factors. In

[3] All CS sections can be found in AICPA *Professional Standards*.
[4] All ET sections can be found in AICPA *Professional Standards*.

such situations, the description of the set of data may need to include the identification of such criteria. In addition, when the characteristics of the data differ from a widely used or expected set of criteria, it may be useful to users to specifically state that fact.

20. Implementation guidance accompanies each criterion. This guidance presents factors to consider when making judgments about the nature and extent of disclosures required by each criterion. The implementation guidance does not address all possible situations; therefore, users should carefully consider the facts and circumstances of the entity and its environment in actual situations when applying the criteria.

21. The description may be presented using various formats, such as narratives, tables, or graphics, or a combination thereof. The degree of detail to be included in the description generally is a matter of judgment unless there are contractual, legal, or regulatory requirements that specify the detail to be included in the description.

Consistency of a Set of Data With Its Description

22. A set of data is consistent with its description if

- each member of the set of data is appropriately included in the population of the events or instances represented by the set of data.
- no events or instances that should be included in the population are omitted from the set of data.
- the elements of each member of a set of data is complete, accurate, valid, and current based on the description.

23. The following table presents the criterion for describing and evaluating the integrity of a set of data. The implementation guidance in the right column of the following table presents factors to consider when applying the criteria. The implementation guidance does not address all possible situations; therefore, careful consideration of the specific facts and circumstances when applying the criteria would be advised.

Criteria	Implementation Guidance
DI1: The description of the set of data includes the intended use of the data.	Determining the data to include in a set of data requires many decisions about the fitness of the set of data for its purpose and intended use. The persons making the determination need to identify the intended users of the data and how those persons are intended to use the data. Failure to identify the intended users and intended use of the data can result in the data being used for a purpose for which the data is not relevant or not valid, complete, accurate, or current for a particular use. Similarly, users of data need to determine whether their planned use of the data aligns with its intended use. If the planned and intended uses of the data are not aligned, the user may make erroneous or insufficiently informed decisions. Examples of intended uses of data include the following: *a.* *U.S. census data*. An intended use of the data from the U.S. census conducted every 10 years is to support the U.S. Congress in determining the apportionment of the House of Representatives. *b.* *Automobile dealer inventories*. An intended use of inventory data from an automobile dealership is to help the automobile manufacturer make decisions about production volumes and pricing and marketing strategies.
DI2: The description of the set of data includes the following: *a*. The population of events or instances included in the data *b*. The nature of each element (field) of the data (that is, the event or instance to which the data element relates) *c*. The sources of the data *d*. The units of measurement of data elements	Whether creating or using a set of data, it is necessary to understand what is and is not included in the set, as well as what elements of the data are included (for example, a field within a record). This information is provided in a definition of the data contained in the set. An inappropriate definition may result in material errors, either due to a material error arising from a single member of a set of data or a material error arising from the aggregation of numerous immaterial errors in the data, including the following: *a.* Incomplete identification of events or instances *b.* Erroneously included events or instances *c.* Inaccuracy in the measurement or recording of data elements *d.* Data that is not fit for its intended purpose and use The definition may accompany the data or be available separately. For simple sets of data, the definition may be implicit in the presentation of data, for example, the column headings of an on-screen television guide lists channels, times, and programs.

e. The accuracy, correctness, or precision of measurement *f*. The uncertainty or confidence interval inherent in each data element and in the population of those elements *g*. The time periods over which the data was measured or the period of time during which the events to which the data relate occurred *h*. The factors in addition to date or period of time that determined the inclusion or exclusion of items in the data elements and population	However, for most sets of data, a formal description is necessary to use the data appropriately. *Identification of the Population* The population consists of the events or instances included in the set of data. Users generally need to understand the factors that determined the inclusion or exclusion of an event or instance from the set, which are typically communicated in the identification of the population. For example, the date of sale as defined by generally accepted accounting principles (GAAP) determines whether a transaction is included in the sales journal for a particular period of time. For the 2010 U.S. census data, the data population includes the households that responded to the census form or that were interviewed by the census takers, over a specified period of time. There may be situations in which the set of data is known to be incomplete or has missing members of the population (for example, records may be missing for a particular date due to the failure of computer storage or an error in recovering data after a system disruption) or because of incomplete elements for members of the set of data. When the set of data is incomplete, the identification of the population should include information on the missing or incomplete members of the population to permit the users to understand the nature of those members of the set of data and permit those users to evaluate whether the data is sufficient for their purposes. For example, a missing temperature reading from a thermometer measuring hourly air temperature may not be material for a set of data used to determine the average daily low temperature for the year. However, the loss of wind speed measurement during a hurricane may significantly affect the usefulness of a set of data used for establishing building codes in a hurricane prone area. *Nature of the Elements* Each member of a set of data comprises elements that are the characteristics of the data that relate to a particular event or instance that has been recorded. In a traditional database, the member of a set of data is usually represented by a record, and the elements are recorded in the fields for that record. Each recorded characteristic is an element. Types of characteristics are measured consistently between events or instances, and the shared nature of the characteristics is described and associated with the elements in which they are recorded. Whether recording characteristics or using the recorded data, the definition of each element needs to be

understood in order to reduce the risk of misstatement of the characteristic.

For example, “attendance at a ball game” might mean tickets sold, persons passing through the turnstiles, or all persons at the venue who were not playing the game or were otherwise employed in game-related activities or serving game patrons. The attendance reported will vary depending on the nature of the different definitions of *attendance*.

Sources of the Data

Understanding the source of the data is necessary in order to reduce the risk of misinterpreting the data during its use. For example, oral and tympanic measures of body temperature yield different results; therefore, a researcher using such data needs to understand its source. In addition, understanding the source of the data may affect a user’s evaluation of the quality of the data. For example, data from independent sources, such as commercial data providers, may be more unbiased or more reliable than data produced by the entity in some circumstances.

Sources of data should be identified at a level of specificity that would permit the data user to obtain the same or similar data given the appropriate circumstances or a user to understand the characteristics of the source of the data.

In addition to understanding the source of the data, in some instances, it may be useful to provide information on how the data was collected and other information that can help the user evaluate the sufficiency of the completeness of the data. Such information may include collection methodology and limitations of the methodology, biases in the collection process, if any, and unusual circumstances in the collection of the data.

Units of Measurement

Measurement of most elements requires the use of a unit of measurement. Because many elements have alternative units of measurement (for example, length measured in meters or feet), identification of the unit of measurement for each element is necessary in order to avoid misunderstanding and potentially erroneous results. For example, in 1999, the Mars Climate Orbiter failed when one piece of software provided data in a different unit of measurement than was expected by the software that used the data.

Unless obvious from the nature of the element, the unit of measurement should be identified.

Accuracy, Correctness, or Precision of Measurement

The measurement of many characteristics has a limit to its accuracy, correctness, or precision. For example, a thermometer that measures with a precision of two degrees may be sufficient for cooking but not for controlling a chemical reaction. In order to collect data accurately, the collector of the data needs to understand the level of precision required, and in order to use the data appropriately, the user may need to understand the level of precision used in the collection of the data.

Uncertainty in the Data and Population

Many types of data have characteristics that are uncertain at the time of measurement and will only be determined at a later time. Until the uncertainty is resolved, an estimate may be recorded based on the underlying characteristics. For example, a meteorologist may forecast rain at a certain location on a certain date. However, providing the measure of uncertainty (for example, the chance of rain is 80 percent based on the historical data for the conditions measured) allows a user to make a more informed decision.

There is no prescribed method for communicating uncertainty. Examples of such methods include the following:

a. Providing the standard deviation of the element

b. Providing information on historical variations

c. Reporting the margin of error associated with polling data

d. Describing a range of possible values for the elements

In addition to describing the uncertainty, it may be useful to describe the persons or the qualifications of the persons determining the recorded value of an element that exhibits uncertainty. For example, it may be useful for the user to know that a financial estimate was prepared by the actuarial department of an insurance company.

Date of Measurement or Period of Occurrence

Identification of the date of measurement of an element or population, or the period of time over which events occurred, is critical to both the measurement and use of data.

Other Factors

The characteristics of data vary depending on the nature of events or instances that form the subject population of the data. In addition, the characteristics of data may vary due to variation or changes in

	measurement. Consequently, the definition of the set of data may require inclusion of information regarding other factors in addition to those specified here. Examples of such characteristics include ownership, classification for security and privacy purposes, access privileges, version, retention and disposal requirements, lineage and audit trail information, and assurance-related information. They also may include changes in the consistency of measurement taken over a period of time. For example, the thermometer used to record hourly temperatures may be replaced by a newer model, resulting in a more precise reading and the elimination of a bias in the older thermometer. For some types of elements or sets of data, there may be specific, defined criteria that are relevant in the determination of the population, the nature of each element, the source of the data, units of measurement, accuracy, correctness or precision of measurement, uncertainty or confidence in the elements, or other factors. In such situations, the description of the set of data may need to include the identification of such criteria. For example, a set of data containing an extract of the general ledger of an entity would usually state that the determination of the existence and occurrence, completeness, accuracy and valuation, and rights and obligations are based on GAAP. In addition, when the characteristics of the data differ from widely used or an expected set of criteria, it may be useful to users to specifically state that fact.
DI3: The description of the set of data is complete and accurate.	For the set of data to be useful, the description of the data provided must be valid, complete, accurate, and current. Failure to include information likely to be relevant to the decisions of the intended users in the description or misrepresentations in the description may result in an erroneous or insufficiently informed decision by the user of the set of data.
DI4: The description of the set of data identifies any information that has not been included within the data or description but is necessary to understand each data element and the population in a manner consistent with its intended purpose.	Whether creating or using data, a significant amount of metadata is needed. The data description is a key source of metadata, but it is usually not practical to include all the information needed to use the data in the data description. For example, the information required to use a company's financial statements includes knowledge of GAAP. When additional information beyond the data description is needed to use the data properly and the need for such information cannot be presumed to exist on the part of the user, the additional information should be identified. For example, in a set of data about crude oil inventories, the description may need to refer to the API

	(American Petroleum Institute) gravity definitions of each grade held.
DI5: The data description is provided with the set of data or is otherwise available to users of the set of data.	Users need the data description in order to properly use the data for processing, presentation, and decision making. Users who do not have such information available to them are at risk of misusing the data, which may result in erroneous or insufficiently informed decisions. Management should provide the description with the data or communicate to users where the description can be obtained.
DI6: The data in the set of data is consistent with its description.	The data within the set is valid, complete, accurate, and current.

Effective Date

24. The criteria are effective when issued.

Glossary

accurate. Subject matter prepared in accordance with criteria is free from error or sufficiently precise. This includes concepts such as correctness and precision of measurement or estimation as well as consistency of representation over time and across items.

complete. Subject matter prepared in accordance with criteria does not omit relevant factors that could reasonably be expected to affect decisions of the intended users made on the basis of that subject matter. This includes whether all members that should be included are included in the set of data as well as the completeness of elements within each member.

criteria. The benchmarks used to measure or evaluate the subject matter.

current. The subject matter is current if the present reality or condition of the subject matter is being represented. This is evaluated relative to the time period or cut-off date of the information relative to its intended purpose and the timing of its use.

data mining. The practice of examining large databases in order to generate new information.

format. The form in which the data is presented, which includes structured and unstructured representations.

information asymmetry. Present whenever one party possesses greater material knowledge than the other party.

member of a set of data. Data regarding a particular event or instance that is included in a set of data.

metadata. A set of data that describes and gives further detail about other data. This includes the appropriate context required to understand the information.

set of data. A defined collection of data regarding events or instances that share common characteristics or relationships.

valid. The data and the elements of the data represent what they purport to represent.