**INCITS Big Data Ad Hoc BD-<tba>**

**ISO/IEC JTC1/SGBD N<tba>**

Date: <proposal date>

**INCITS Big Data Ad Hoc**

**ISO/IEC JTC 1/Study Group on Big Data**

|  |  |
| --- | --- |
| **DOCUMENT TYPE** | Change Proposal |
| **TITLE** | Add Veracity as a Big Data Characteristics |
| **SOURCE** | Mark Underwood, Krypton Brothers |
| **PROJECT NUMBER** |  |
| **STATUS** | Proposed |
| **REFERENCES** |  |
| **ACTION ID.** |  |
| **REQUESTED ACTION** |  |
| **DUE DATE** |  |
| **Number of Pages File Size (KB) Diskette Number Issue Number Filenames** |  |
| **LANGUAGE USED** | English |
| **DISTRIBUTION** | P & L Members  SC Chair & Secretariat  WG Conveners and Secretaries |

**REFERENCES**

[JTC1SGBD N0015] “Draft SGBD Report to JTC 1 v1.0”, April 25, 2014

# Introduction and Rationale for Proposed Addition

In a presentation made at the San Diego meeting, I argued for Provenance as a major concern of Big Data standards organization. I am proposing Veracity as the fourth V in the Big Data V’s, and suggest that veracity is a useful near-synonym for provenance. Whether veracity/provenance is a subset of metadata, or vice versa, is a fruitful discussion worth having. In the meantime, these seven reasons are recommended as part of the rationale for the inclusion:

1. In the absence of provenance information, data can be incorrect while secured.
2. Privacy concerns continue to be the primary interest of the public in Big Data. The legacy of big data providers in credit scoring (Experian, TransUnion, etc.) has led to mistrust in the ability of individuals to own, and to correct data about themselves. This concern is not addressed in the other V’s. It is a problem not addressed by simply securing incorrect data. Without making veracity a top level concern, the Big Data standards activity is likely to be perceived as self-serving geek-speak that ignores these legitimate concerns.
3. Because re-identification, complex event processing, digital forensics and data fusion are activities that already ingest Big Data, the potential for misattribution, error and faulty inference-making will increase.
4. The Internet of Things will add a huge overlay of machine-to-machine data for which automated systems will rely. Establishing provenance for each individual sensor – especially firmware/software-based sensors – is domain-specific; for instance, it could require nontrivial calibration that is configuration-aware and time-critical. In addition, such provenance data is potentially a Big Data source itself.
5. Big Data systems must be designed with user interfaces that take into account human factors aspects of sensor data. For instance, some sensors may be critical in whose absence automated systems should be halted. In other settings, such as dashboards, different sensors can be substituted. The process of degrading a Big Data system that requires on high volume, high velocity data streams could well entail intelligent use of specific metadata and provenance elements.
6. Building resilient systems, or understanding the feasibility of an approach to resilience, requires an integrated approach to data source quality, reliability. Big Data systems that rely on aggregate data must take into account any limitations of those sources. Data availability and timeliness (e.g., information on demand), for instance, may be a feature of provenance. For example, some data sources may be available only at scheduled times, or only when dependent systems are healthy, or even upon power or weather conditions.
7. GRC issues may apply that are not captured elsewhere. For instance, data could be rented and cannot be used after the rental period. Royalty payments may be derived based upon provenance data elements. Data may be legally licensed for use only in certain countries, or only in limited contexts. Data could be subject to court ordered retention and formatted regulations. The other V’s do not capture any of this.

An IBM-sponsored study Big Data found that lack of confidence in data sources is a major concern. As [summarized](http://on.wsj.com/1tHKW97) by the *Wall Street Journal*,

A fragmented approach can result in a breakdown of trust among different groups of people who may be accessing, interpreting, and using data in different ways. This gap stems from a basic distrust about who is qualified to competently analyze and act upon the data. A lack of trust among executives, analysts and data managers can significantly impact the willingness to share data, rely on insights and work together to deliver value. IBM's study found that a trust gap among individuals is a leading indicator of lack of trust in the veracity of data. When this happens, the overall costs to a business are high.

A number of other sources recommend including Veracity in the definition of Big Data:

* Inside Big Data <http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>
* Patricia Saporito (SAP blogger) <http://blogs.sap.com/innovation/big-data/2-more-big-data-vs-value-and-veracity-01242817>
* Objectivity.com <http://www.objectivity.com/news/blog/gartners-missing-vs-value-and-veracity/#.U1>
* Big Data for Dummies Cheat Sheet <http://www.dummies.com/how-to/content/defining-big-data-volume-velocity-and-variety.html>

While Veracity is important for existing data sources, it is extra special important for Big Data because…

I therefore propose to add Veracity as an additional V in the Big Data definitions in [JTC1SGBD N0015]

# Proposal

In [JTC1SGBD N0015], add a new section following Section 5.3.4 “Variability”:

5.3.5 Veracity

Veracity is a unifying principle for a set of related quality management and governance processes not exposed elsewhere in the Big Data canon. Understanding the Veracity of data is extremely important to many applications. It addresses a critical Big Data concern of the lay public, which is how to ensure the accuracy and usage life cycle for personally identifiable information.

Big Data Veracity potentially encompasses one or more of the following:

1. The provenance of data and the life cycle of processes that created it. For example, “synthetic” data can be created for segmentation purposes, and data such as gender could be inferred from genetic markers or self-report on survey instruments.
2. Information about data accuracy or calibration
3. Legal constraints upon data distribution or usage, such as national boundaries or court subpoenas.
4. Legal or regulatory limitations on the types of joins permitted with other Big Data sources.
5. Time-constrained aspects of data, such as population measurements superseded after a specified period of time, or data which cannot be used until a minimum period of time has passed.
6. Metadata regarding boundaries of usefulness for data; for instance, outlier values which are to be disregarded, which may reflect sensor or other varieties of failure, or which are intended to trigger resilience assurance measures.
7. Data which cannot be used for specified purposes until one or more GRC events have taken place.
8. An associated ontology for units of measure. For instance, in 2014, the National Science Foundation cited the absence of uniform meanings for units of measure as a key limitation in exploitation of Big Data for science and engineering.
9. Operational, fee-for-usage and metering metadata
10. Portals to support audit and forensics. For some systems housing PII this can mean standing up portals to display how a single individual’s data has been used and to allow procedures for data correction.

Veracity is not an absolute measure. Data that has sufficient quality and provenance for one use may not be sufficient for another use.