The Life Cycle of Trusted Data:

From Acquisition to Persistence… Or Not

Jay Hollingsworth

CTO

Energistics

Abstract

Continuing from last year's paper on the importance of trust in knowledge work, the categories of trustworthiness, and how apparent data quality impacts perceptions of trustworthiness, this paper carries the trust concepts through into analytics and data management.

We begin by identifying facets of trustworthiness which are mathematically related to data uncertainty and introduce paths to quantification of those values. These are things like obtaining the catalog drift values to use on uncalibrated sensors or finding r-squared in the original papers defining common computations using empirical correlations (like Archie's Law for water saturation).

We then discuss how to conveniently represent trustworthiness as metadata and carry that information along as data makes its way through the seismic to simulation (or exploration to abandonment or whatever a company uses) workflow.

We end with a discussion of the data management aspects of trustworthiness. Does trustworthiness become stale over time as data ages? How would we persist trustworthiness scores in a database? Do we need to persist them or are they derivable into the future?

Introduction

Knowledge workers of all kinds – geoscientists and engineers, data scientists, analysts, planners, managers, etc. – are uncomfortable using data they do not trust. They know the quality of their work or the decisions they make will be adversely affected by bad or inappropriate data. These knowledge workers will spend as much time as they feel is necessary examining and “correcting” the data they need to use to perform their work. Factors influencing how long this process will take include their individual feeling about the sources of information, prior bad experiences with use of improper data, and the riskiness and visibility of the result of their work product.

Establishing or recovering a level of trust in the official sources of enterprise data is a key – though often unwritten – objective of corporate data management efforts. Many internal IT projects use the word “trusted” in their titles. Some commercial applications which focus on creating trusted data use a variant of “trust” in their product name. These are all recognition of the value of trust in the data chain.

In a previous paper[1], we established the fact that trust is a different concept from data quality, and that in fact people trust the use of poor quality data where they explicitly understand the level of trustworthiness and can assess whether their use of imperfect data is appropriate for their task; they simply take the uncertainty into account.

In the prior paper we also established the finite number of dimensions of trust in E&P data and showed that each could be explicitly measured.

In this paper we examine more closely those dimensions which relate directly to classical uncertainty measures and consider the impact on data management practices in supporting data scientists and other knowledge workers as they consider which data to use for a given task (or as they try to improve the quality of the data at hand).

[1] From Data Transfer to Data Assurance: Trusted Data is the Key for Good Decisions, 22nd PNEC Conference & Exhibition, Jay Hollingsworth *et al.*

Dimensions of Trust

In our prior work we elected to use definitions of trustworthiness put forth by IBM[2] in 2011 and added to them. We use this framework to show that a data transfer standard could be used to convey the components of trustworthiness from one system to another so a user could establish their own level of trust in a data source. There are additional details – facets – along the dimensions which we will not repeat here.

There are three dimensions of trust:

1. Trust in the Data Source
   The qualitative reputation and a quantification of the reliability of a source of data inherits to the data retrieved from that source.

2. Data Lineage
   Data lineage creates trust in a data by tracing its provenance from its origin through the history of its processing.

3. Data Security
   This dimension encompasses trust in the security of the underlying systems and (in the case of real-time data) security of the transmission mechanism between the sensor and the user.

Some authors have created schemes to try to assign scores which relate to the trustworthiness of a particular piece of data or of data sources. Most of this work relates to information of a more qualitative nature – trust in government or financial institutions, for example. These scores are derived from assigning values to the various facets which exist within the three dimensions above (see the referenced papers for more details).

In upstream oil & gas the data we are concerned with are more likely to be the result of an activity which measures or describes something. This data could be automatically acquired or manually entered, but regardless of how they come into the system they share common aspects of gaining and maintaining users' trust and in requiring proper data management.

[2] The Third International Conference on Advances in Databases, Knowledge, and Data Applications, Trusted Data in IBM's Master Data Management, Przemyslaw Pawluk *et al.*

Measurements

In order to apply proper data management processes to measurements, it's essential to have a clear definition of what a measurement is. Psychologists have developed clear rules regarding what true measurements are – since so much of what they try to measure isn't truly quantitative or maybe even measurable at all. A good if somewhat academic definition is[3]:

> *Measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules*

The cited article goes on to define several categories of rules a measurement might follow but the one of interest to us is when the numeral is a ratio against a standard. Meaning, there is a standard which exists somewhere and the measurement I am making is, say, 12.6 of those. For example, there is a metal bar (located in Sevres, outside Paris) which serves as a standard meter; so a length measurement is really comparing some desired length to the length of that bar.

Modern measurements concern themselves with their accuracy and uncertainty, and any measurement apparatus will have some kind of traceability back to that standard, even if that path is no longer known.

Trust in a measurement according to this definition depends on the user of the data being aware of the process which has been followed in acquiring a value and deciding whether that process has resulted in a value suitable to the use at hand.

The Process of Measurement

The process of measurement in E&P has several steps, each of which introduces its own uncertainties and possibilities for error. Data managers need to decide now many of these steps they will present to their users to establish and retain their trust in the official source of truth:

- Identification of a desired measurement, followed by
- Realizing that it can't actually be measured, but that something related can be
- Selection and acquisition of a sensor to measure that related thing
- Placement of the sensor
- Calibration of the sensor
- Taking of a reading
- Conversion of the analog reading to a digital value
- Calculation of the desired measurement from the actual one

[3] Science, vol. 103 no. 2684, June 7, 1946, page 677, S. S. Stevens

Identification of a desired measurement

This step is often overlooked in data management circles, but why are we asking for a particular measurement? To what purpose will it be put? This knowledge dictates what data needs to be acquired.

If a company is optimizing the drilling of a particular well they need downhole measurements for things they can control like weight on bit, the rotational rate of the bit, the flowrate of mud through the nozzles on the bit, the physical properties of the mud, etc. They also need measurements like rate of penetration and the inside diameter of the hole to make sure they are drilling a quality hole as quickly as they can.

In a producing well the engineers need to know how much a well is flowing and at what pressure.

Geoscientists want to know the characteristics of the rocks away from a wellbore and the properties of the fluids in those rocks so they can safely place wells at the best location.

Unfortunately none of these things is directly measurable.

There is no human at the bottom of well who can directly observe the conditions downhole. The only rock properties we know with any certainty are for the rocks we removed from the well (by drilling or coring). You simply can’t directly measure flow inside a closed pipe.

There are things which are directly measurable – lengths can be measured against a measuring tape, fluids can be poured into a calibrated vessel, weights can be balanced against standard counterweights, etc., but most of the measurements of use in oil & gas are computed indirectly.

What We Can Measure

If the user of data is lucky, the data they want can be directly measured and they can trust the data they need.

In the usual case the values they need must be computed from other measurements. For any given desired value there is a balance between the accuracy required for a particular use and the expense of taking real measurements. The relationship between the measurements taken and the desired value can fall into several familiar types of relationships - empirical, stochastic, heuristic, deterministic and possibly others not detailed here.

The nature of the equation to be used and its associated uncertainties can be found from literature and will contribute to the trustworthiness of the resulting data.

For example, in a production scenario everyone needs to know how much a well is flowing, but as mentioned earlier one cannot directly measure flow. One can listen to the sound made by fluid as it flows through pipe, there could be a propeller inserted into the pipe and the rate of spinning would be proportional to the flowrate, a heated wire could be placed in the fluid and the rate of cooling would indicate the flow, and there are many other techniques.

The most common method is to measure pressures before and after a hole of known size in a pipe of known size plus to measure the temperatures (for calculating the fluid density) and to otherwise already know the gas composition.

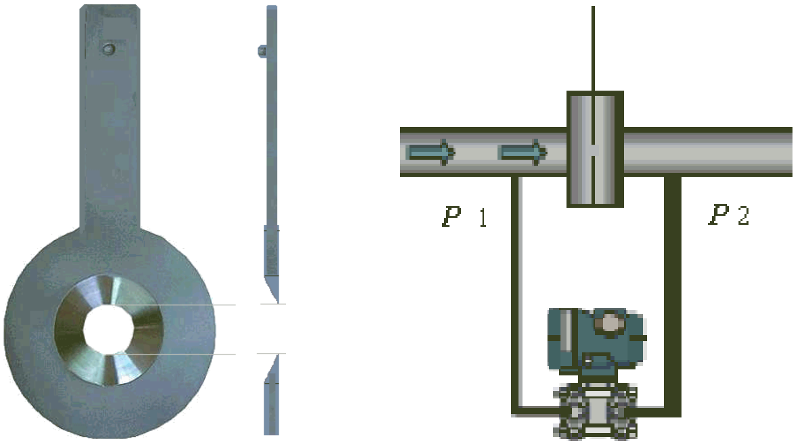


Image by Alfaomega at German Wikipedia

From these measurements there is an industry standard method of computing the mass flowrate (per ISO 5167):

$$q_m = \frac{C}{\sqrt{1-\beta^4}}\,\epsilon\,\frac{\pi}{4}\,d^2\,\sqrt{2\,\rho_1\Delta p}$$

The rationale by which this method was chosen, the uncertainties associated with this empirical relation and the methods by which the input data are acquired all become part of the dataset which needs to be managed for the knowledge worker.

Other examples will be discussed in later sections.

Selection and acquisition of a sensor

Clearly, once the desired measurements are chosen sensors need to be placed in contact with the “objects or events” so the measurements can be taken.

Again, there is a balance between the expense and durability of a sensor and the accuracy required of the sensor. There may be some uses of sensor data which do not require extreme accuracy – the gas gauge in a car doesn’t need to be perfectly accurate – when it gets close to E you stop and get more gas. For real-time control of processes on a rig or in a plant, more accuracy may or may not be required. For improved accuracy a measurement contractor may synthesize results from redundant sensors. The dependency of any real-time process on the quality of the data feeding it is a topic of current research[4].

Another consideration is the operating range of a sensor – either the range of the variable being measured or of external factors. Many pressure sensors are accurate within a certain pressure range

[4] SPE-181076, *A Step By Step Approach to Improving Data Quality in Drilling Operations: Field Trials in North America*, Pradeep Ashok *et al.*

but far less accurate at very low (close to atmospheric) or very high pressures. All sensors will have a temperature range in which they are accurate. These will influence the choice of a sensor as well.

The data user may have little or no influence on the precise sensor selected for a particular data acquisition but they may be able to specify the accuracy of a measurement service.

Placement of the Sensor

The next point of uncertainty is the location of the placement of the sensor.

At a minimum the user of the data needs to know where the sensor is located. Clearly there are many cases where the sensor is placed some distance from the desired measurement and corrections have to be applied. But there are many cases where the placement of a sensor affects the reading itself.

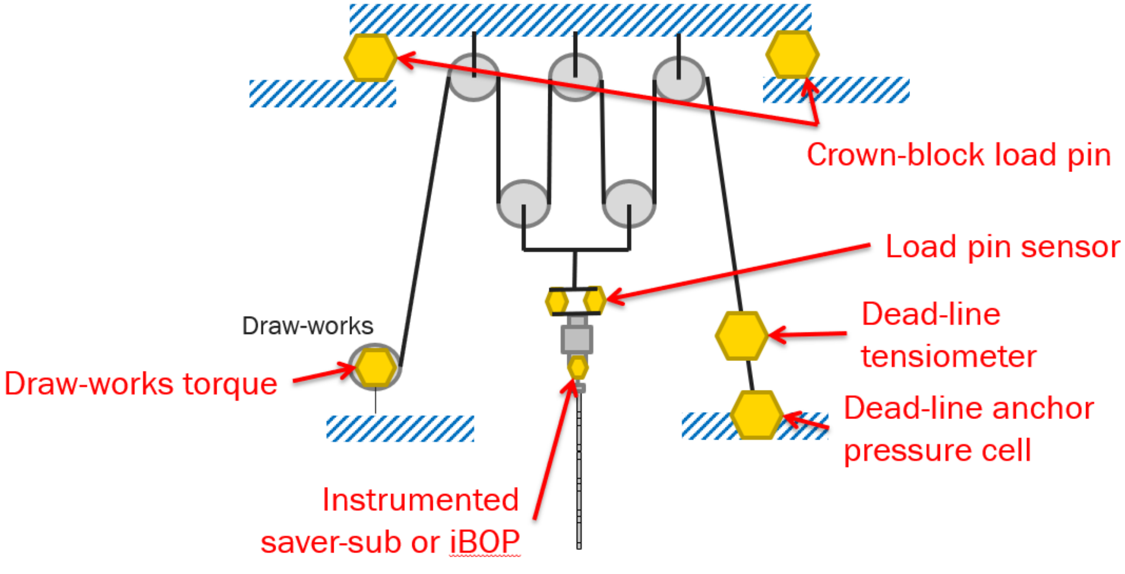
MWD and LWD tools are included in the bottomhole assembly but their results are generally only used until wireline surveying and logging can take place – the presence of the rest of the BHA and the drilling conditions negatively affect the accuracy of the results. Using surface pressure gauges for drillstem testing means that the bottomhole conditions cannot be truly calculated (because the drillpipe causes difficult-to-calculate pressure drops in multiphase flow).

Difficulties in knowing precise measured depths are well known and are a calculation. So any downhole sensor always has uncertainty in its depth.

In addition, there are many cases where the same channel may come from multiple sensors at different locations on the rig. Depending on the rig state (is it in slips, or pulling out of hole or drilling ahead or picked up off bottom and circulating or whatever) or the accuracy sought, the user may choose a reading from one of another location.

For example, the downward force experienced at the drill bit – the weight on bit – cannot easily be directly measured. As a surrogate, the weight is calculated by subtracting the upward force exerted by the hoisting system from the total weight of the drillstring as suspended in the well.

In this case (sketch courtesy of the International Research Institute of Stavanger) there are 6 possible locations to measure this upward force. Some are accurate, others less so due to friction and other forces.



Finally, it’s important to realize that the location becomes part of the primary identifier of the sensor itself. If there are two load pin sensors on the traveling block then one may be the north one and the other the south one. Or if there are three tension meters there is an upper, a middle and a lower one.

Data managers need to maintain this location information “metadata”.

Calibration of the Sensor

Everyone understands that a sensor needs to be calibrated in order to be accurate. And that as soon as the calibration process has ended and the sensor put back into service it will continuously become less accurate. This process is called sensor drift.

What is not as well understood is that this sensor drift can itself be calculated and the resulting uncertainty quantified. The value for sensor drift can come from two sources:

1. The initial manufacturer’s specifications should have a value for drift. This will be for the sensor model, not for the specific sensor installed. If data managers make this catalog information available, analysts can estimate the uncertainty due to poor calibration.
2. The other way to quantify the sensor drift is through repeated calibration. If the calibration activities are recorded and the observed error prior to recalibration is noted, the drift over time can be computed. This is not possible unless information from the calibration activities is preserved.

The other piece of information required to make use of sensor drift is a knowledge of the time elapsed since the previous calibration activity. If one knows that a temperature sensor drifts at 0.1 degrees Fahrenheit per month and it's been 5 months since it was calibrated then 0.5 degrees can simply be added to the nominal accuracy of the sensor.

It's worth noting that field experience indicates that some kinds of sensors get better as they age – they "mature" – while others get worse through wear and tear. Tracking the sensor drift over time will help quantify this effect as well.

It should also be noted that there is a difference between zeroing a sensor and calibrating it. Zeroing is like resetting a bathroom scale when one steps off it; it still may read high or low but zeroing it will help. A zeroing activity should also be noted and preserved alongside a true calibration.

The calibration information may be located in a maintenance management system or it may be in a specialized calibration database. There are details about calibration – traceability to an ISO or NIST standard, the procedure used, the identity of the calibration technician – specialists may be interested in.

Taking of a Measurement Reading

Data people are generally comfortable with the notion of keeping the actual measured value and know the importance of preserving its original unit of measure. But there is much more "metadata" available which should be preserved.

Remembering that measurement is an activity, then any information one would preserve about any generic activity is appropriate for measurements.

Who measured it when using which apparatus or sensor?

What were the ambient conditions? Is the temperature within the nominal operating conditions for the sensor? How long has it been since the sensor was calibrated (if this is not otherwise available from a maintenance management system)?

The actual raw measurement may be buried in a combined sensor or transmitter box and may not be easily available; the transmitter will typically output a voltage, a current in a 4-20 mA control loop or (in older equipment) a 3-15 psi pneumatic pressure. Regardless, if this kind of raw reading is stored a knowledge of the raw value and the full-scale settings of the sensor and its deviation from linearity are required to make sense of it.

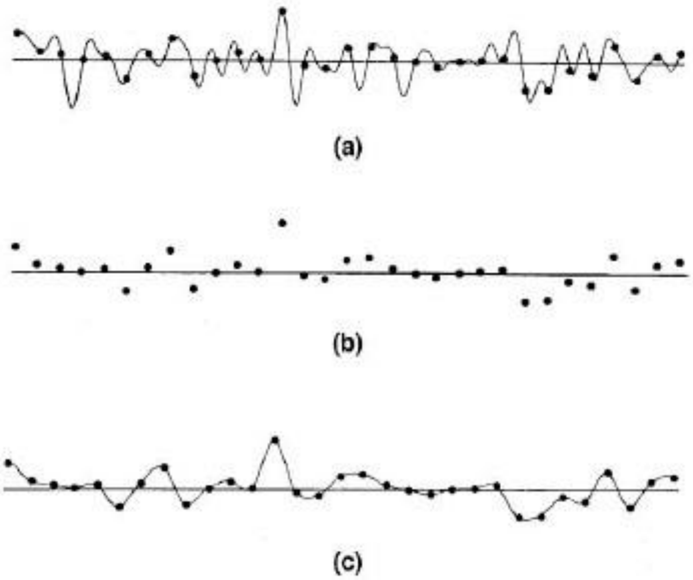
Finally, it should be noted that this raw voltage or current is always an analog, continuous reading from the lower limit of the sensor's range to the upper limit, just like looking at a car's gas gauge from empty to full. And just like in your car, you need metadata to understand the relative reading – meaning you need to know that your car's gas tank holds 30 gallons to understand that if it's half full then there are 15 gallons left.

Conversion of the Analog Reading to a Digital Value

Once there is an analog reading available, the value needs to be converted to a digital number to make it useful to computer analytics. This analog-to-digital (A/D) conversion may be explicitly visible because it takes place in a separate piece of equipment or could be hidden inside what looks like a digital sensor.

The analog to digital conversion process is always an approximation and needs to be calibrated in the same sense that a sensor is calibrated. Digital sensors - which are typically a combination of a sensor, a transmitter and A/D converter or a sensor directly connected to an A/D converter – may have a single calibration process which covers both the sensor and the A/D conversion.

The A/D conversion “discretizes” a continuous reading signal along two dimensions. In time, the signal is divided into individual readings. This can be thought of as a sample rate – samples per second or a rate in kHz – or may be viewed as the time interval between samples as is done in seismic data – ½ ms is a common seismic data sample interval. The process of sampling is always lossy along the x (time) axis and will lose all information whose frequency is higher than half the sample rate (the Nyquist frequency).[5]



In addition to the time axis, the discretization of a signal also happens on the y (amplitude) axis. When converting a continuous analog signal to a digital one, it is converted into a binary number along its full-scale range. If a single 8-bit byte is used to represent a full-scale range value, there are only 256 possible numbers to carry it. That means if a sensor has a 4-20 mA full-scale range each milliamp of current could only be represented by 16 possible numbers. So if, for example, a 0-100 psi pressure sensor had a 1-Byte A/D converter, the digital reading could only be accurate to about 0.4 psi[6] regardless of how accurate the sensor itself was. Carrying additional decimal places of precision is pointless.

Presenting the sampling rate and amplitude resolution will improve trustworthiness in the data because the error band can at least be calculated if it can’t be avoided. Information about the A/D conversion may be relevant in high-accuracy calculations and should be available to knowledge workers in case they need it.

[5] Image from the SEG Wiki https://wiki.seg.org/images/5/52/Ch01_fig1-5.png

[6] 100/256 to be precise

Calculation of the Desired Measurement from the Actual One

The final step along the road to creating and managing trustworthy measured data closes the loop back to the beginning. We wanted to measure something, but we found we had to measure something different which could serve as a starting point from which the desired number could be calculated. Now we need to compute that desired value.

At the beginning this journey, we decided which equation or algorithm we were going to use so we would be sure and have the measurements we needed to reach our goal. Documenting the choice of algorithm and preserving what software (and which version) was used to execute it is just the beginning of the metadata needed to judge the trustworthiness of the result.

Based on the measurements, we needed to choose parameters for our equation or maybe those parameters change depending on the geology or sea state or ambient temperature or other external factor, so we need to preserve those values as well.

The parameters themselves will have some uncertainty and the dependency of the final result on that variability is also really important to be able to document. This depends on the mathematics involved and is likely described in the original published paper (plus its errata two issues later). Making this available to a knowledge worker will improve their trust in the whole process.

For example, if a reservoir engineer wants to estimate the oil in place in a reservoir, he or she needs to know the total volume of rock, the portion of the rock which contains fluid (the porosity) and the locations of the gas/oil and oil/water contacts. Unfortunately, finding the oil/water contact isn't as simple as one might hope – it isn't a simple perfect boundary; the water saturation changes over some vertical distance as the oil layer transitions into the water aquifer. Generally, for this kind of simple estimate one chooses where the water saturation is at 50% as the location of the boundary. So we need to measure the oil/water boundary (or the depth at which the water saturation is 50%). Since we can't directly measure either one, we can just find a way to compute the water saturation along a drilled wellbore and just pick the 50% point. Geologists typically use Archie's Equation to do this.

$$\mathrm{S_w}^{\mathrm{n}} = \frac{\mathrm{R_w}}{(\Phi^{\mathrm{m}} \times \mathrm{R_t})}$$

where

R_w (water resistivity) - computed from SP log - need R_{mf}, T_{mfr}, etc.
R_t (formation resistivity) - from a deep resistivity log like RILD
n - saturation exponent - varies from 1.8 to 4; 2 is a guess
m - cementation exponent - varies from 1 to 3; 2 is a guess
Φ - porosity - from neutron porosity like NPHI or sonic like SPHI

In this example, we need to preserve the values of the assumed saturation and cementation exponents, the parameters from the logs used - R_{mf}, T_{mfr}, etc. – and the measured logs used. This equation will produce a new computed log channel representing water saturation (one value per input log reading), which also needs to be preserved. Then the final value will be the measured depth at which the water saturation is 50%, which also must be kept.

Data Transfer Standard Support of These Aspects of Uncertainty

The latest generation of Energistics standards – WITSML v2.0, PRODML v2.0 and RESQML v2.0.1 – all fully support the metadata needed to describe these aspects.

Data Assurance Object

The DataAssurance object is common across all of the Energistics standards.

It declares whether each bit of data contained in a related standard object fails a pre-defined corporate data policy. To save bandwidth the Data Assurance object isn't normally carried for data which conforms to the policy, since that is expected to be the norm.

A Data Policy is:

> One or more rules that must be applied to assess the data. Conformance to a policy would classify the data as Trusted

So a Policy is just a named set of rules which trustworthy data should follow. An exception to a policy is flagged as a failure along with the identities of the rules within the policy which failed.

The rules may be determined by the values of the data itself, or may be defined externally.

For example, a company could have a policy called the "Well Location Adequacy Policy." Within this policy could be three rules:

1. Every well location must be re-surveyed with 30 days of completion of the well
2. Every well location must be given in WGS84 as a latitude, longitude pair
3. Every well location must be in a map projection as a northing, easting pair

In a case where all three rules are satisfied there would be no Data Assurance object. In the event that it has been two months and the location hasn't been re-surveyed yet, any transfer of the well location data would carry a Data Assurance object which would note that the data fails the corporate Well Location Adequacy Policy because a new survey is not present.

The Data Assurance Object is in the shared portion of the new suite of Energistics standards, so it is available to be used by all of them and for workflows that require a combination of Energistics standards.

Activity Model

Another new feature common to all the standards is the Activity model.

The Activity objects carry knowledge that an activity was performed (and who did it and when using what software) which may have consumed data, may have used certain parameters and may have produced some data. These could be real E&P field activities like perforating or calibrating a meter, or could be a calculation activity like starting up a reservoir simulator.

For example I could have an activity which uses Archie's equation, the neutron porosity channel called NPHI1, the resistivity channel called LLD2 and cementation and saturation factors of 2.0 to compute the water saturation curve WS3.

The Activity objects can be carried along with any WITSML object, just like Data Assurance or can be transported independently.

Trusted Data Via Data Assurance in Energistics Standards

The key point about the Data Assurance process embodied in the new generation of Energistics' standards is that the data – flawed or not – is always transferred. There is just some extra descriptive metadata carried along as needed to describe the processes used to create or treat the data.

It should be clear that the Activity model and the Data Assurance object taken together satisfy all the facets of trust defined before and could describe the aspects of uncertainty presented here.

The Data Lineage and Data Security facets are covered by use of the activity model. The Trust of Data Sources group is satisfied by defining appropriate rules within policies.

Data Management Summary

The objective of this discussion has been to describe sources of uncertainty in measured data and to show that making the details of these aspects of the measurement process available to knowledge workers will give them faith that the data they are using are suitable to the use they will put them.

Knowledge workers know how to incorporate uncertainty into their work and with this information they will be able to quantify those variabilities which are germane to the task at hand.

All of the data and metadata described in this paper are valuable for analytics, but completely satisfying the desire for complete transparency and availability may be unreasonable in a single data environment.

Some of these data are essential:

- Raw measurements – they are what the company paid for so need to be preserved
- Parameters on those raw readings are needed to make sense of them
- Some information about the measurement activities like when and where they took place
- Manufacturers and model numbers of the measurement equipment
- Computed values on which decisions were based – Sarbanes-Oxley and other audits, contracts

Additional data from less-familiar sources may be brought together to give additional trustworthiness:

- Maintenance management systems – calibration activities
- Rigsite data acquisition & control systems – OPC-DA/UA via WITSML
- Equipment specs – actual (not nominal) sizes, sensor drift & location
- Procurement – what was ordered, what was received, dates of services