From Data Transfer to Data Assurance: Trusted Data is the Key for Good Decisions

By Jay Hollingsworth and Jana Schey, Energistics

Abstract

When thinking about data transfer standards, the traditional view is this: get data from point A to point B in a standard format that can be consumed by various applications, with minimal transformation required. But the truth is, what companies really need is data they can trust as early in the process as possible to make decisions that improve operational efficiency, enable worker safety, and lower costs.

As the industry has worked with Energistics to develop second-generation data transfer standards, data management challenges and use cases were front and center in the effort to ensure our standards can help companies meet operational objectives. Initial use cases focused on drilling workflows, but it quickly became evident that ensuring trusted data using a data assurance capability should be expanded to support any workflow throughout integrated operations.

To address data assurance requirements and consequently increase users' trust in their well-related data, WITSML v2.0 supports data assurance policies and rules that describe the fitness for purpose of any data being transferred. This capability then allows the effective inclusion of information such as, sensor data precision, calibration, and many other parameters. This schema enhancement enables the transfer of this information between applications that validate data according to an organization's data assurance policies and rules.

These enhancements were developed alongside and are released concurrently with the new Energistics Transfer Protocol (ETP), which enables significantly faster data transfer.

WITSML v2.0 does not determine data quality, but it supports the ability to provide assurance that the data are auditable, traceable, and fit for purpose. These initiatives support organizations' efforts to standardize data management, simplify data gathering for data analytics, and help reduce software development life cycles.

The second generation of the Energistics family of standards (WITSML, RESQML and PRODML which are based on a common technical architecture) all support this data assurance capability. As a result, organizations are beginning to operationalize data assurance and governance using Energistics standards. This paper discusses how data management objectives are enabled by Energistics data transfer standards with WITSML as the example.

Introduction

In the November 4, 1991 issue of the Oil & Gas Journal, an article published by Chevron's chief geophysicist spawned countless reorganizations within corporate IT groups and justified the costs of practically every data management project over the following 20 years. The nugget of information in the article that kicked off all this activity was the statement that geoscientists spend 60% of their time "looking for data".

This statement has been widely misinterpreted as a literal statement that geoscientists—and by extension all technical staff in oil and gas companies—spend three days of every week for their entire careers searching for data. In the '90s, presumably this meant rummaging around in dark archives filled with crumbling boxes full of dusty well and seismic files and in offices abandoned by their laid-off former occupants. In the '00s the staff would have been prowling around in electronic document management systems and the project databases that had proliferated underneath their Unix-based technical computing applications. In more recent times, one assumes they were hoping Alta Vista or Google would illuminate the path to their data.

In fact, the statement is not an expression of what geoscientists do; it's about what they don't do. Geoscientists of the day—and technical staff up to the present day—feel they only spend around 40% of their time performing the work they set out to do and in attending meetings and so on. The other 60% of their time is their equivalent of a driller's non-productive time—the hours a drilling crew spends waiting on supplies, like cement, to be delivered or conducting various kinds of tests or safety drills or otherwise not spent with the drill bit on bottom making hole. This non-productive time was simply labelled as "looking for data" for lack of any better name for it. It is indeed true that technical professionals spend a certain amount of time honestly performing queries and looking for information that is not readily available. This is particularly true for specialists whose data isn't clearly identified as such (like sequence stratigraphy or basin analysis or rheology). Most of the non-productive time is actually spent in preparing an adequate TRUSTED data set, so they can properly begin their work.

Trusted Data

Do not trust all men, but trust men of worth; the former course is silly, the latter a mark of prudence. – Democritus

Trust is a "firm belief in the reliability, truth, or ability of someone or something" (Oxford University Press 2017). We trust our loved ones. We trust our pets not to suddenly return to their wild nature and attack us. We trust the other drivers on the road to behave rationally and not intentionally cause an accident.

We develop trust in sources of information as we grow up. We learn to rely on our senses, even though we know they can be fooled. We trust our parents and teachers in school to give us the information we need to move through life. We trust measurement devices like rulers and tape measures and bathroom scales and kitchen cups and spoons to give reasonably accurate readings. We trust computers to do as we ask and to give us the information we request.

Trust in anything can be broken. A bite from a neighbor's dog would cause us to no longer trust him, even though his owner may assure us that he's normally not like that and he might never bite us again. A visit to the doctor's (presumably more accurate) scale might cause us to no longer trust the one in our own bathroom.

A bad result at our job could cause us to lose trust in the official source of truth we are supposed to use to make workplace decisions.

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 are adversely affected by bad or inappropriate or improperly licensed data. These knowledge workers will spend as much time as is necessary to reach the point that they can trust the starting point of a particular item of work. This could be weeks or months of time expended in examining and "correcting" the data they use to perform their work, depending on 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.

In extreme circumstances, users with poor experiences with use of the "official" data in a company may create their own trusted data store; in the data warehousing world these are called "rogue data marts." These are typically made by extracting a copy of the "official" data from wherever it resides and then altering or adding to the data in ways the user believes improve it. The rogue data mart then becomes the trusted source of data for this particular worker. In many cases, co-workers may approach the owner of the rogue data mart and ask to make use of it, so that it becomes the trusted data source for an increasing body of knowledge workers, whether or not the data are truly better for any given purpose than the official source of truth.

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 that focus on creating trusted data use a variant of "trust" in their product names. These are all recognition of the value of trust in the data chain.

Dimensions of Trust

Several academic sources regarding the quantification of trust were consulted for use in this paper, and we have elected to begin with the definitions of trustworthiness put forth by IBM in 2011 (Pawluk et al.) and to add to them. The authors of that paper were creating a scoring system to attempt to quantify trustworthiness; in this paper we use their framework to see whether a data transfer standard could be used to convey the components of trustworthiness from one system to another so that a user could judge their own level of trust in the data source.

The IBM authors' classification includes three main dimensions (data lineage, data security and trust of data source), which are explained below. We use those dimensions and their categories and add a couple of categories of our own, as described.

Data lineage

Data lineage creates trust in a source of data by tracing it from its origin through the history of its processing, which includes the sub-categories listed here.

Origination

Origination identifies the original source of the data. For manually entered data, this is the identity of the user who created the data, whether manually entered or the result of a human interpretation. For measured data, this could be an identifier of the sensor that measured the value.

Traceability

Traceability is a kind of extension of the origination facet that describes the various storage locations and processing history of a single piece of data. Traceability is particularly important in E&P data because so many data types are corrected using various algorithms.

If a well location has been de-projected from its original system to an older geodetic system, subsequently transformed to a newer system (like WGS84) and then re-projected, those facts and the parameters used are very important to accurate positioning.

If a well header has been composed of data from a mixture of sources, being able to see which source was used for each value—because those choices may change from one entry to another—adds to the trustworthiness of the data, even if the user disagrees with the choice of source; the transparency itself improves the users trust in the data.

Stewardship Status

Stewardship in this context is about knowing whether the data has been through manual steps like data re-entry, or whether the data processes have been handled by automated systems.

To the original IBM categories for data lineage we add:

Originating Software Inputs and Parameters

Much E&P data originates as a result of calculations based on multiple inputs, which do not easily fall into one of the prior categories. Trust may be based on the name and version of the software that was used to produce a particular piece of data, the algorithm used to compute it, and the various inputs.

For example, one of the most basic pieces of real-time drilling data is the average downhole weight on bit (WOBA). This is always a calculated value and may be computed using one of many different algorithms, which could have taken as input one or more sensors on the

rig and will have been averaged over an arbitrary time or depth interval. Understanding these factors is clearly key to having trust that the data is suitable to the use to which it will be put.

Data Security

This dimension of the IBM paper 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.

Authentication

Authentication identifies to a data user whether the data they are considering using comes from a system that has access limited to specifically identified users, or whether it comes from a source that could have been tampered with or in which someone may have easily impersonated a trusted source.

For real-time data, a system with strong encryption gives the user confidence that the data hasn't been intercepted and altered along its travel from the wellsite to the office (a man-in-the-middle attack).

Authorization

Authorization is a companion to authentication and discloses the process by which users are granted access to a data source. For example, if all that is required to gain access to a data source is a process of self-registration (like in Twitter or Facebook) then data from that source would be less trustworthy than data that comes from a source where authorization is granted by a more rigorous third party.

Roles Policy

A rules policy conveys whether the source system limits the roles that authenticated users can play. That is, can everyone allowed in the system alter or load data, or is each user granted a specific set of roles for each of the datasets to which they have access? Finer-grained control of who is allowed to edit which data improves the trustworthiness of the data.

Auditing Policy

An auditing policy describes whether or not the data source can "remember" who made changes to the data and when. This is critical in cases where sensitive data may have been erroneously altered or loaded improperly. The simple presence of such an auditing process increases trustworthiness of the information in the data source.

To the original IBM data security categories we add:

Licensing policy

This policy describes the uses to which the data may legally be put over a defined time period and whether the appropriate licensing fees have been paid. Users will likely not trust data obtained illegally.

Disclosure Policy

This category covers the circumstances under which a piece of data may be disclosed to a partner, to the public, to a regulator, or to anyone at all in the employee's company. This could be a date after which a piece of data may be made public. Conformance to this policy ensures that a user does not accidentally release "tight-hole" information because they didn't realize the data was supposed to be kept secret.

Again, knowing that the information they are using is not identified as being "tight-hole" data improves the user's trust that the data is appropriate to their use.

Trust of Data Sources

The IBM paper uses this dimension to collect the remaining categories of trust that are not related to lineage or data security.

Believability

Believability covers the degree to which a data source can be considered a credible source of truth. In this category, the user wants to know whether the individual piece of data fails to conform to any of the rules established by the company that comprise a policy regarding the truthfulness of the data. This might include a data range validity check, the calibration schedule of measurement instruments, completeness, a starting data value for a month matching the closing value for the prior month, etc.

This topic is developed further later in this paper, but the mere presence of a system that conveys measures of the believability of a piece of data dramatically increases the perception of trustworthiness, even for specific data that have been flagged as failing to conform to one or more rules.

Reputation

Reputation is a subjective measure of the user community's confidence in the trustworthiness of a data source. The reputation of a data source is enhanced by an awareness

that it is covered by a formal data governance process, with data owners identified and actively working on data quality projects.

The reputation of a data source may also be improved if it is identified by a reliable authority as being the "official" place to go for this kind of information. If experience proves that the data is actually unreliable, however, then the "official" stamp of approval may do little to improve its reputation.

Objectivity

Objectivity is not normally a concern among technical data information sources, but certain kinds of contextual data and much traditional business data (like customer records) may be subject to the personal bias of a single individual or of a community. In an environment where such a bias may be carried over into an analytical environment, the ability to carry a label indicating so would improve the trust a user has in the data; they can choose to use the data or not.

Reliability

Reliability is a measure of whether the data source typically provides the correct information or not, and the circumstances under which the data values may be erroneous.

For example, if it is known that a particular sensor is very accurate when the ambient temperature is below 100 degF and that its accuracy declines rapidly above that, being able to convey that fact (and possibly the current temperature along with it) gives the user the ability to decide whether or not to rely on that particular sensor reading.

WITSML™ v2.0 and Data Quality

WITSML v2.0

The Energistics community recently published WITSML v2.0, a new generation of its ubiquitous drilling data transfer standard. This new version of the standard was six years in development and represents a break from the past in its use of newer Web technologies to move data faster and with less delay from the field, while retaining aspects of the prior generation, which preserves much of the existing investment companies have made by continuing to use XML for contextual data and keeping much of the data model intact.

WITSML v2.0 is based on what Energistics calls its Common Technical Architecture (CTA), which is an updated set of technologies that enable seamless transfers involving any combination of all three of its standards: WITSML™ for downhole data (including drilling data), PRODML™ for surface and production data, and RESQML™ for earth model and reservoir simulator data. This CTA includes, among other things, a set of common data types and reference values which are shared by all of its standards.

The requirements-gathering process lasted several years and was driven by documentation of use cases that would justify changes made to existing XML schemas and any new work that would be needed. Some clusters of these use cases spawned separate workgroups to document them and to support the model work that would satisfy them.

One of the groups of high-priority use cases was around data quality.

Data Quality Use Cases

Sixteen use cases were originally grouped together under the data quality umbrella, and these were reduced to three high-priority ones and one late addition, all of which were addressed in WITSML v2.0:

W-V2-R007 **Title:** WITSML2 support for data validation functionality

Summary: Many Sensors have a range of valid values. The idea is to implement features in the standard that would detect data entries violating the rules/valid values, and make it easy for clients/applications to notify users. NB: No filtering of data.

W-V2-R006 **Title:** WITSML2 support for data quality in real time

Summary: Ability to correct data live and propagate the correction or a warning in the stores applications that have downloaded the original data.

W-V2-R003 **Title:** WITSML2 support for data validity flag in real time

Summary: Ability to correct data live and propagate the correction or a warning in the stores/applications that have downloaded the original data and to propagate the process used so it could be unwound later.

NEW **Title:** WITSML2 support for data auditability and traceability

Summary: When applicable, the end user would like the ability to optionally receive auditability and/or traceability data generated from the point of origin to the end user.

Data Quality Becomes Data Assurance

During the development of the data quality capabilities of WITSML v2.0, it became clear to the group that the phrase "data quality" had too many concurrent meanings for it to be useful as a concept that could be practically implemented in the context of a data transfer. The phrase typically refers to a project undertaken to provide a scoring of a database based on questions of completeness (are all the data we need there?), consistency (is there a latitude wherever we have a longitude?), correctness (are all the depths positive going downward), etc. and then a subsequent project to correct the problems.

A data transfer process—as supported by a data transfer standard—doesn't score values in a database and can't fix data while it's moving.

What is needed in a data transfer is the capability of carrying enough metadata so that the user can trust that the data is good enough for the purpose to which it will be put. Then the user can decide.

This process is called 'data assurance', and the group renamed itself and the WITSML model as such.

Data Assurance in WITSML v2.0

Data assurance was defined by the group as: a governance process that applies policies to data providing assurance that the data is fit for purpose, auditable, and traceable. Data assurance supports real-time, near real-time and historical analytics.

This process is supported by a simple Data Assurance data structure and a more complex Activity structure, which can be carried along with any WITSML v2.0 object.

Data Assurance Object

The simple Data Assurance object declares whether each bit of data contained in its companion WITSML 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, because that is the expected 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 that trustworthy data should follow. An exception to a policy is flagged as a failure, along with the identities of the rules of the policy that 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." This policy could have three rules:

1. Every well location must be re-surveyed within 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, no Data Assurance object is transferred. If two months has passed and the location hasn't been re-surveyed yet, any transfer of the well location data would carry a Data Assurance object that notes that the data violates the corporate Well Location Adequacy Policy because a new survey is not present.

The Data Assurance object is part of the Energistics CTA, so it is available to be used by all Energistics domain standards.

Activity Model

Another data model in the new shared common area is the Activity Model. The Activity objects carry knowledge that an activity was performed (and who did it, when, using what software) which may have consumed data, may have used certain parameters, and may have produced some data. These activities 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 that 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 v2.0 object, just like Data Assurance, or can be transported independently.

Trusted Data Using Data Assurance in WITSML v2.0

The key point about the data assurance process is that the data—flawed or not—is always transferred. There is just some extra descriptive metadata carried along as needed.

It should be clear that the Activity Model and the Data Assurance object taken together satisfy all the facets of trust defined above. 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.

There could also be a company rule that says that a pressure sensor must be calibrated every six months, but at the end of that time, the pressures and volumes calculated from that sensor could be flagged that the sensor may be unreliable. The user then knows the questionable reliability of the data and can use it where that level of accuracy is appropriate—to know that the well was flowing at all, for example.

Summary

We have shown that trusted data, enabled by technologies that underpin data governance policies and business rules, enables sound business and operational decisions.

The project to produce WITSML v2.0 developed the concept of data assurance and delivered a standard that supports the concepts needed to maintain trusted data delivery in an enterprise.

Although WITSML v2.0 does not determine data quality, it does enable the ability to provide assurance that the data are auditable, traceable, and fit for purpose. This capability supports processes and rules that standardize data management, simplify data gathering for data analytics, and help reduce software development life cycles.

These capabilities also provide a foundation for individual companies to establish data governance policies, roles and business rules helping to ensure trusted data that enables safe, reliable, and compliant operations.

Energistics SIGs are now exploring the potential to create a standard blueprint that will help companies leverage and realize the full potential of these capabilities.

Credits

We would like to thank the members of the WITSML Special Interest Group and the participants in the Data Assurance group within the SIG for their efforts in supporting trusted data delivery.

References

Oxford University Press. 2017. Oxford Living Dictionaries (English). https://en.oxforddictionaries.com/definition/trust (accessed March 2017).

Pawluk, P. et al. 2011. Trusted Data in IBM's Master Data Management. Presented at The Third International Conference on Advances in Databases, Knowledge, and Data Applications. Held in St. Maarten, The Netherlands, Antilles, 23-28 January (proceedings: www.thinkmind.org/download_full.php?instance=dbkda+2011) (accessed March 2017)