JCMB TECHNOLOGY

Data for the smart grid

Data Quality Assessment Process

JCMB TECHNOLOGY TECHNICAL SERIES

Smart grid readiness program

195 St-Francois-Xavier, Delson, QC
Phone 450.632.5844 • Fax 450.632.3207

Table of Contents

Summary

The smart grid is probably the most difficult demand placed on electric utility operators ever. This green, self-healing and magical grid's starting point is the infrastructure we have now. How ready are we?

It is no secret that the all inclusive smart grid is going to be a challenge to implement. Its definition is only starting to emerge from the theoretical vision. Although there is a lot of talk about what it has to be, very little consideration is given to the actual state of the art in many utilities, both large and small.

In many cases the current infrastructure is far from smart. There is much reliance on ephemeral knowledge for its resilience. Even if you could replace every experienced electric operator by a set of intelligent switches tomorrow, what will those switches do without an accurate knowledge of the network that is to be switched?

An experienced operator can make sense of network model and data errors and make appropriate decisions in times of need. Intelligent devices however do not have this ability, they are quite inflexible and fault intolerant with regards to the data they use to make decisions.

This document explores the requirements for a solid smart grid data foundation and provides a road map to it.

The smart grid

So much has been written with regards to the definition of what a smart grid should do that we all have a more or less clear view of what the future of this rejuvenated grid is like. What do we need to make it happen though? One thing for sure is that the SMART part of this grid will require some thought.

We learned from our past implementations of distribution information systems such as GIS and OMS that good and accurate data is the foundation upon which successful systems are built. Data is a deceptively simple word but it can hide many complex concepts:

Knowing what information is available

Knowing what is missing.

Understanding the structure in which the information is kept

Understanding the relationships that exist between those data objects

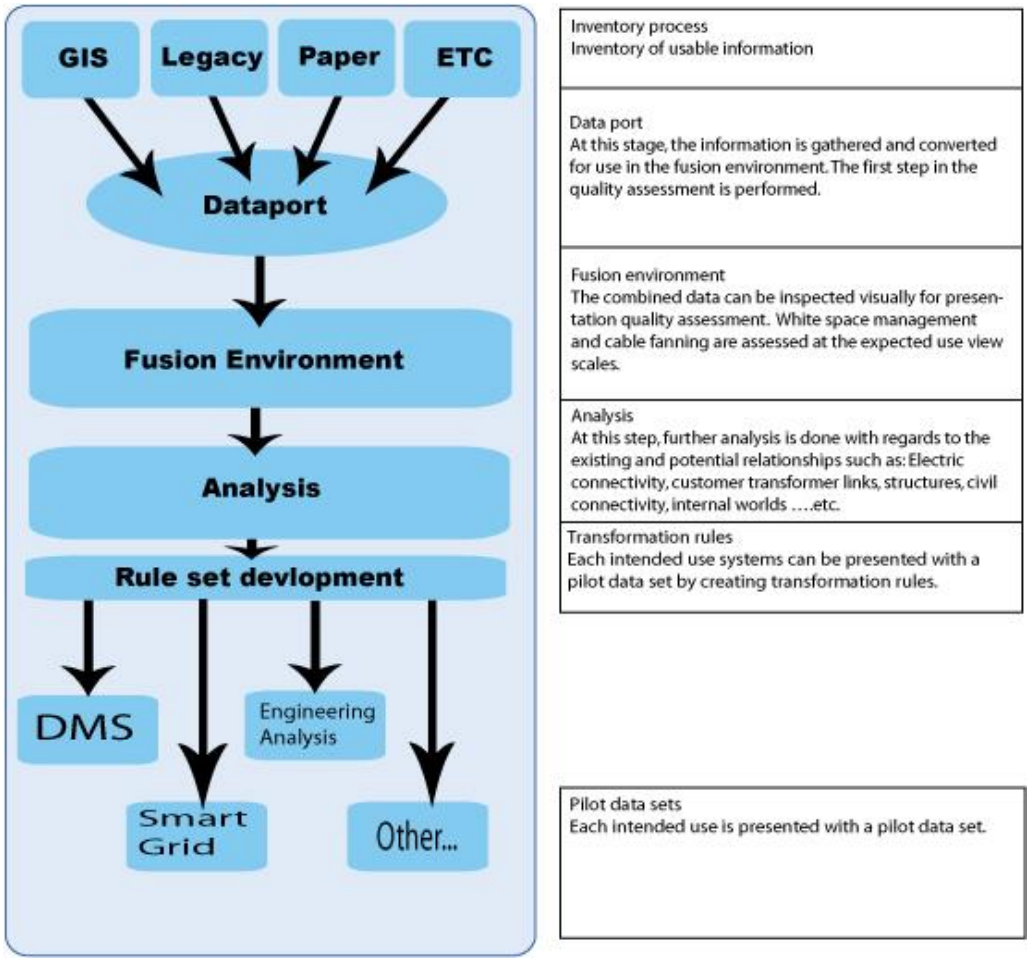
And finally, the integrity of it all

This complexity makes the scope of the data requirements for the smart grid is far reaching in consequences. Intimate knowledge of the current state of the legacy with regards to data is especially important in the early stages of our smart grid initiatives. Successful implementations often depend on this careful analysis.

JCMB has created a structured method for gathering this information, the DQAP process.

DQAP overview

DQAP stands for Data Quality Assessment Process, a process by which intimate knowledge of a utility's data assets is methodically acquired. It is an accurate picture in time of the quality of the data and the infrastructure that holds it. This process is essential to create an accurate assessment of the usability of this data in creating the applications that drive the smart grid.



DQAP is a structured process that can be executed in phases. Each phase is autonomous, has its own set of reports and serves as the foundation for the next.

The process starts with discovery, a process in which data and possible sources of data are inventoried and classified. Data is then processed in the Dataport to create a source data environment in Fusion. With this data set, a test suite is created to validate the usability of the existing information. This test suite leads to the creation of an inventory

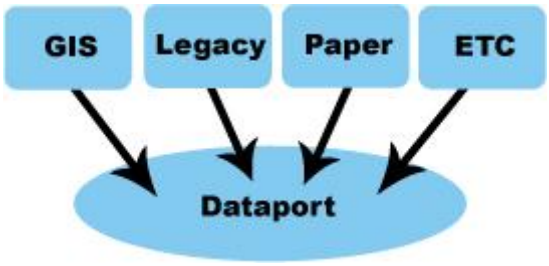
of defects and voids that exists along side the data legacy. This step creates a very accurate picture of the current data situation.

The Fusion environment also provides the opportunity of visualizing the data in new and meaningful ways for detecting possible use problems such as overlap, unreadable elements at specific view scale values, cable fanning etc. When evaluating the data for use in a new system, the knowledge gained through discovery can be compared to the data requirements of the new system. In the case of a smart grid initiative, those could be the data requirements of a Distribution Management System for example. With those requirements in hand and the first phase inventory knowledge acquired through the Fusion environment, a delta report is prepared. This report highlights the differences between the actual data and the data assets required to successfully implement the new system.

All along the DQAP process, the results of the analysis is documented in the DCSM. DCSM stands for Data Certification Standard Manual. This manual is the specification that is prepared in light of the analysis of the first and second phases of the DQAP. It contains the specifications for the data correction process needed to bridge the gap between the actual state of the data discovered herein and the requirements of the new system(s) being implemented.

After this analysis stage, enough knowledge and experience has been gained with the data in DQAP to create test data samples. This is an opportunity to test and gain practical experience of the new data knowledge by applying it to a small sample set of the original data. It transforms the first phase theory into practical real world, hands on, knowledge. Our vast experience of data preparation has shown this practical, 'Hands On', experience to be the only sure way of accurately predicting the effort required for the data collection and correction phases in a project. Knowledge and experience is the key for reducing implementation risk for all parties involved in the implementation of new system initiatives, especially with regards to the smart grid.

The discovery process



Data assets are built over long periods of time. Documentation for their structures or contents is often out of sync with the data itself. Databases with information gathered over decades often have cumulative errors and obscure or obsolete information in them. It is important that the data horizon be explored at the lowest level. The inventory process looks at all the systems that are susceptible to provide useful information.

GIS and other Asset Databases

GIS systems and Asset Databases is often the first data store that is inventoried in the process. Information that is harvested from those systems is multifaceted. They contain discrete information about the facilities but also contain valuable relationships that are often more important than the device data itself.

Unlike the discrete data, relationships can be buried deep into the systems. Modern systems such as Esri and Smallworld have well documented topology engines that can be harvested in a structured manner using our off the shelf interfaces. Legacy systems on the other hand must often be reverse engineered in order to get to their information. Our data harvesting technicians have extensive reverse engineering experience and have extracted data from various types of systems both obsolete commercial offerings and proprietary purpose built.

In any system, we invariably look for ways to build correct relationships:

- Device to Device (Connectivity)
- Device to Structure (Civil relationships)
- Device to Supply point (Customer connections)
- Structure to Structure (Civil connectivity)

Other data types are also harvested from related systems when they can be used in the certification process. For example, data from meter reading routes and reading sequences can be used to validate customer to transformer relationships. Although this data is not used directly, it can be used to enforce logical rules in the certification and error correction process.

Legacy systems

Many legacy systems have been reverse engineered by our data mining technicians to harvest useful data sets. We look at the code base and the data to determine the extent of harvesting that can be done. In some instances, we can use our previous experience with older FM systems based on Framme, Autocad and other Cad like systems. In other cases we have taken data out of home grown systems with diverse back end storage from flat file to Oracle databases.

The process is basically the same when looking at legacy systems. We are looking for the same discrete data and relationships found in more modern GIS systems.

Paper maps

Paper maps can be used as a primary source of information or used to supplement electronic data. In some instances, we will find details with regards to some data relationships in paper form only. The most common example of this is when looking at the internal configuration of underground structures. Electronic forms such as GIS often neglect to model the details of the internal structure or the civil relationships that are required to trace empty conduit sections. The complexity of modeling the networks to this level of details was often beyond the abilities of earlier implementations.

When paper maps are a primary source of information, an inventory of the paper information is created. The project is then treated like a conversion but this does not prevent JCMB from doing a certification based on a definition of the conversion and a precise inventory of the paper information.

Other information

Other related information is inventoried in this phase as well if required:

- Landbase
- Geocode databases
- Aerial photography
- Alternate geo databases
- CIS/Meter data/Meter routes or read sequences
- Asset management
- Mobile computing
- Etc.

The *Dataport* process

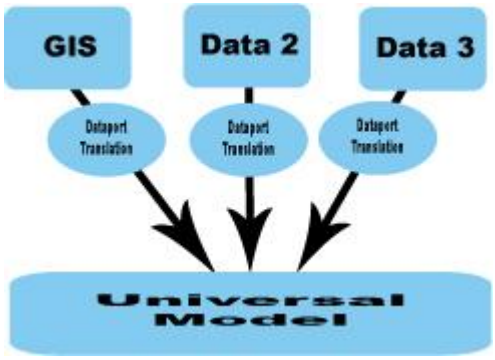
JCMB has created the Dataport technology to facilitate the format conversion of utility data for import or export. This technology has a rule base engine that runs 'Data Enhancement Rules' in the Fusion jargon. The ability to create translation, transformation and enhancement rules in this engine without having to program a new interface is the reason for its high efficiency.

JCMB can often leverage an already made rule set when importing data from often encountered source such as Smallworld, Esri, Oracle and others. JCMB is constantly adding base rule sets to its Dataport technology. When starting from one of those off the shelf base rule set the configuration of the import translation can be done quickly.

Why import data into Fusion

The goal of the DQAP process is to certify a data set for a specific use. Considering the large spectrum of different data elements and relationships involved, a human intellectual or theoretical analysis alone would be hit and miss at best. You need to use technology to supplement the human mind when dealing with such large and complex data sets. Our certification is based on experimentation, not intuition.

Fusion universal model



Data is taken by the Dataport into the Fusion universal model. This is a utility model that contains a super-set of utility data. It has a standard place for everything and even surrogates objects to replace those that are missing in the input data.

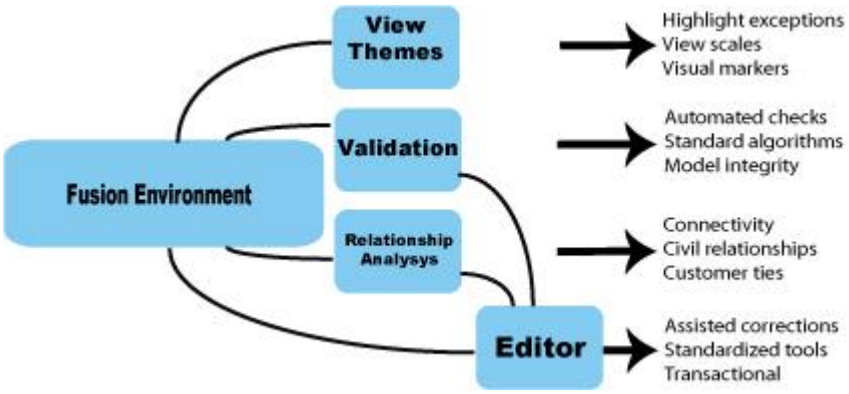
Standardizing data at this level enables JCMB to use off the shelf algorithms for visualization, analysis and validation. It also provide our technician with a constant user interface for applying edits and additions.

In fact, creating the initial Dataport translation rule sets provide actual labor but mostly large time saving when considering the benefits of holding the data set in our universal engine.

- Rich tool set designed especially for the certification of utility data sets
- Off the shelf analysis algorithms.
- Off the shelf validation
- Off the shelf editor.
- No need to train analysis and correction staff.
- Minimal custom code, stable certification code base

Taking data into Fusion leverages a well defined, stable and constant certification process that relies on repeatable technologies.

Technology assisted analysis



Much of the DQAP certification process's accurate assessment of a data set's readiness for the smart grid is founded on Fusion's ability to consistently highlight the data exceptions, making them easy to spot and assess. It is similar to putting a sample under a microscope to examine the invisible structures of its constitution. In such an environment the defects and omissions are amplified. There are three main components to Fusion's functionality.

Thematic view engine.

The ability to highlight specific conditions visually is created by a rule based engine that analyzes the information at the rendering phase when data is displayed in a Fusion environment. The rendering phase occurs when a data set is processed to be presented to the user in a viewer. This is much like a GIS system presents a map on screen. However the similarities end there. Fusion does more than a simple rendering of symbology. Data is intercepted in the Fusion rendering process. All network elements create a rendering that highlights specific features of the object, of a relationship it has with other objects or a combination of those two. A rule set for rendering in Fusion is called a view theme.

Typically, a data set would be rendered in more than one view theme. Each view theme would be tasked to highlight certain features for analysis. Many of these themes are Quality Centric, made to render quality features of the data in an obvious, highly visual way. Examples of themes available in Fusion:

- Connectivity
- Civil connectivity

- Customer supply points
- Underground vs overhead
- Cable size
- View scale specific Overlap
- View scale specific cable fanning
- Geoposition analysis
- More…

New standard themes are being added to the list continually. Specific themes are also developed to highlight customer specific data features.

Validation

There are two levels of error detection. The first involves data integrity mechanisms that are built into the environment. This engine is trigger based and normally prevents erroneous changes in the editor. They also double as an error detection system as the triggers log the errors upon object creation, thus keeping a list of anomalies with the data itself. Those lists are then used to help users locate and correct the data sets.

Example of validations performed in a typical DQAP process:

Electrical	*Data*	*Spatial*	*Model*
▪ Phase ▪ De-energized ▪ Meshed ▪ Meter-Transformer ▪ Connection rules ▪ Propagation rules	▪ Mandatory attributes ▪ Normalized values ▪ Data integrity ▪ Cross validation ▪ Inter-system conflicts	▪ Equipment location ▪ Landbase reference ▪ Proximity rules ▪ Annotation ▪ Overstrike ▪ Visualization ▪ Node reduction	▪ Incompatibilities ▪ Limitations

Relationship Analysis

The Fusion environment contains standardized relationship analysis algorithms that are used to discover, diagnose and correct relation sets such as:

- Primary connectivity
- Secondary connectivity
- Structures
- Civil structures
- Civil connectivity
- Customer to feeding device

These relationship traversal algorithms can be further specialized to detect or fix specific aspects of data requirements.

Like the validation engine, the results of this relationship analysis can be kept in the environment to assist data mining technicians in the exploratory or remedial phases of the DQAP implementation.

The DCSM

DCSM stands for **D**ata **C**ertification **S**tandard **M**anual. This manual is the bible of the certification process. Its creation follows all the DQAP steps, documenting all findings with an eye on the final intended use for the data. This document becomes the source of information used for business analysis. It also serves as the technical specification for the interventions that are required in the implementation phase.

This document contains:

- Compliance requirements for data certification
- The list of all the available data elements (the inventory)
- A description of where the information is
- A description of how to get to the data
- A description of the target
- The delta analysis of the source and target
- A description of the transformation rules from source to target
- A description of data capture rules when required
- Statistical analysis of the information (quantitative)

It is important to note that the DCSM is built along side the experimentation and it can be done in phases. For example, the DQAP process can be started with the inventory and the target system description. This would enable a DQAP process to create a delta analysis for costing purposes at the planning stage of an intervention. The DQAP process can then be resumed if the economics of the projects are favorable.

Economic studies

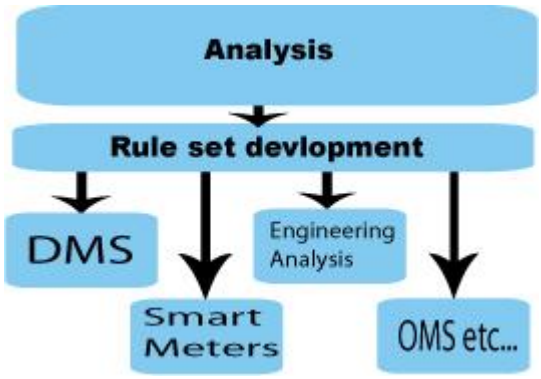
The first phases of the DQAP are well suited to the feasibility phase of new system implementations, especially with regards to smart grid interventions. It provides valuable information on the costs and effort (time) involved with the data aspects of a project.

The information harvested as part of the DQAP process enables very accurate assessment of:

- The state of the data assets
- The cost of evolving the data assets to the quality or completeness levels required by the new systems
- The effort (time) that this will take
- A clear presentation of Data facts for a cost/benefit analysis

Those economic factors can then be used in the overall analysis of the new systems that are to be implemented, especially in a smart grid effort. This knowledge insures that proper funding can be acquired in the early stages of the project.

Gaining knowledge through experience



One of the benefits of the DQAP process beyond economic analysis is the ability to produce target data sets for evaluation, testing or experimentation. Once source data set is imported in the Fusion environment, it can be exported to other systems using the same translation technology that was used to gather it. The Dataport technology that was used on input can be re-configured with a new rule set to produce data in a format that is compatible with the target system.

This approach favors knowledge through experience by providing real data sets to experiment in the target systems. Furthermore, datasets can be produced in more than one format which provides a common data base for evaluating competing technologies.

Simulating missing information

The Dataport technology is rule based. When a rule set is developed to translate data from the internal universal model to a vendor specific structure, it can also be made to contain rules that supplement, simulate or deduct missing information.

For example, input models with underground transformers modeled without the load break elbows can be supplemented with the elbows by the Dataport rules. In fact, this rule is one of many standard rules that is already available in the Dataport.

In conclusion

The DQAP process provides utilities with a repeatable data quality tool, a certification methodology. This process presents an accurate picture of the data assets. Most importantly, it does this through experimentation, creating hands on knowledge in the process.

This structured method for dealing with large data pools is an essential part of planning any new distribution system, especially in the context of a smart grid initiative. It can be used to assess the effort in bringing data to the requirements of the new smarter grid as well as provide an accurate measure of quality in every day data maintenance operations.

In conclusion, the DQAP process provides:

- An accurate assessment of the quality of current data assets.
- An accurate assessment of the effort (time) required to prepare the data legacy for use in the target system (smart grid).
- An accurate assessment of the cost of doing this preparation.
- An accurate base on which to build a cost benefit analysis of the data preparation phase of the initiative.
- An accurate knowledge of what data is missing and the impact this has on the initiative.
- A small data set that can be used to experiment and evaluate target systems with real data.