Paper No.
02-1219

BUSINESS AS USUAL?
CORROSION IN A
CIRCULAR ECONOMY

Applying the Big Data Approach to CP: Opportunities and Pitfalls

Scott McKelvey
Corrosion Service Company Limited
100, 9871-279 Street
Acheson, Alberta
Acheson, Alberta, T7X 6J4
Canada

ABSTRACT

Pipeline integrity is embracing the technologies that have been produced by the Big Data revolution. Database access, machine learning algorithms, and analytics tools are no longer the domain of researchers and IT experts, and can be easily deployed to improve our use of pipeline integrity data.

We will discuss the data governance requirements borrowed from Big Data as applied to CP, a case study involving CP measurements with differing levels of data structure, and rules and lessons learned related to the Big Data approach.

Key words: Cathodic protection, data management, big data, data mining, data integrity

INTRODUCTION

Management of CP systems is all about data: collecting, storing, sorting, analyzing. Improved analysis methods are available to us resulting from the application of Big Data technologies and approaches. Experience has shown that the actual methods for managing CP data tends to be left up to individual users, who will invariably perform these tasks in inconsistent manners. This limits the usefulness of CP data beyond providing evidence of compliance inspections, as the effort required to “clean” data is often more than the effort spent doing actual analysis.

There is a desire to tie CP performance indicators (e.g. polarized potentials and rectifier output current) to less abstract, *physical* pipeline integrity performance indicators (e.g. ILI metal loss features and external ER corrosion probe readings). One-off engineering analyses of this manner have been performed before, however to do this on a large scale, across thousands of miles of pipe on dozens of pipeline systems, with data inputs from a variety of sources, a Big Data approach to managing CP data presents the only realistic and practicable solution to accomplish this goal.

This paper will discuss data governance requirements to increase usefulness and reduce management labour for CP data, using a case study where both poorly-structured and well-structured CP data were analyzed to correlate CP performance to external corrosion. Lessons learned from automated CP data analysis, and guidelines for implementing good CP data governance are also presented in this paper.

[*Author's Note: The author of this paper is by no means trained a data scientist or expert. However, their experience in the CP industry has identified the need for at least some awareness of data science for anyone who owns CP data.*]

THE BIG DATA APPROACH

"Big Data" is a recent term to describe technologies used to manage and analyze very large (i.e. gigabytes or terabytes) sets of data in a productive, efficient manner. Data mining and machine learning are two examples of Big Data in action. CP data in its scope and scale is quite modest by comparison, however Big Data technologies can still be applied to CP data to provide faster and more meaningful data analysis.

There are data governance requirements around these large data sets, but these can be applied to CP data all the same.

BIG DATA PRE-REQUISITE: DATA GOVERNANCE

Before considering the actual implementation of Big Data technologies, the end goal of this implementation must be clearly defined. In the author's case, the end goal was to run well-defined analyses on CP datasets for a variety of assets a typical pipeline system. For example, "*I have five years' worth of daily readings for 100 rectifiers. Each reading is time stamped, and includes DC volts and DC amps. Let's use this dataset to trend groundbed resistance and identify those that are failing.*"

Experience has shown that the level of effort required to do this sort of broad analysis scales exponentially with the quantity of devices or assets being analyzed; generally, the same type of information is expressed differently depending on the data source. Rectifier outputs, for example, could come from a multitude of sources:

- Pipeline maintenance crews / electricians
- In-house CP technicians
- CP survey contractors
- Remote monitoring devices

Each of these sources have their own idiosyncratic data formats, storage locations, and level of data reliability.

This is just one example of the issues that arise on a system of even modest complexity without strong data governance, but in most cases, more effort is spent finding data and rearranging it than on the actual analysis.

Implementation of data governance to reduce this effort, through policies and specific technologies, is straight forward; the key is to actually do it, and to do it consistently. Recalling the original objective to perform analyses on CP datasets, the data governance rules must ensure that the CP datasets are:

1. Uniform, and rigid enough that most users would be unable to break the uniformity
2. Scalable
3. Easily accessible by multiple users
4. Maintainable

Each of rules are essential. Let's discuss each in turn.

Data Uniformity and Rigidity

Excel is the de facto standard “database” tool used in engineering in general. The main attractions are its ubiquity and its perceived versatility; virtually all business PCs come with it, most technical staff have some familiarity with it, and it seems up to the task of storing data at first glance.

Our dataset needs to be rigid enough that it maintains uniformity between individual readings, such that they can all be treated the same by any later analysis. If each user is permitted to measure and enter data their own way, this will inevitably result in non-uniform data. For example, one of our data sources listed earlier might report rectifier volts in units of mV, or another user might use a different unit, and enter it as a text string (e.g. “P/S OFF = -1.1 volts”). They might also decide to include a new column to track some other reading type, or new rows to track new measurement locations, and before long the “template” Excel workbook has a life of its own.

Excel does have some built-in methods for forcing data types, and for “locking down” cells or sheets, however this has proven [N.B. to the author, at least] to be a band-aid solution for inadequate user training, and results in a less user-friendly tool.

Purpose-built database tools are better at mandating this sort of uniformity. For example, an attribute for a rectifier reading could be a numerical input called “VOLTS_DC,” and input value ranges could be set to it to attempt to ensure mostly valid data is input. The rules around this particular input can be changed en masse by the database administrator if updates are required.

This task is simple to accomplish in a proper database system, however in Excel, this would require format updates to each cell, column, or worksheet that contained a “volts DC” value. It might be possible to create this structure in Excel up front, but is impractical to maintain over time [N.B. the author has failed in the attempt].

Scalability

Scalability suggests a large amount of data from multiple sources; in the CP milieu, it means a lot of similar data from a multitude of different users, measurement equipment brands, for an ever-changing list of assets, and updated continuously, in perpetuity.

The primary limitation to Excel in this regard is that it scales to large data sets poorly. Beyond a file size of a few MB, Excel is frustratingly slow and unstable to use. If a single file is used to house many individual data sets, our example of using it to store daily rectifier outputs is only practical for a handful of rectifiers before the file itself is unwieldly. Splitting the data out into multiple files amplifies the data uniformity problem discussed previously.

Using a dedicated database tool, that is designed specifically to scale well for large data sets, is the recommended approach. Historically, database tools such as Access were not user friendly enough for general use. Modern databases tools, such as any SQL-based system, and user-friendly front-ends, such as Sharepoint, are simple to create, update, and maintain by “layman” technical users [N.B. the author, for example]. There is enough basic capability available out of the box, with room for expandability by more experienced users.

By simple virtue of using a modern database, that is designed with gigabytes of data storage in mind, there will be no issues with scalability when used to manage CP data [N.B. you’ll be long dead before you run out of storage for CP data].

Easily Accessible by Multiple Users

Following on the previous two points, the dataset must be accessible by multiple users (or even other databases, such as a remote monitoring web database) if it is to grow to any meaningful size.

Attempts to enforce a single Excel sheet that is updated by multiple users has proven to be impractical without significant manual oversight. Multiple files, each with a filename that attempts to encode the user, date, and the other flags are frequently the result (e.g. *"Scott's CP mesurmints_last Wednesday_Final_2")*. Worse still, sharing these files through emails is not a practical solution for more than a handful users.

Remotely accessible databases, with user-specific login credentials, address this problem, and are generally out-of-the-box features for modern databases and easily deployed. Most modern databases also have web interfaces, removing the need for dedicated licenses to be installed on computers before use; this is especially useful given the abuse to which field laptops are subjected, and opens up the possibility of using smart phones or tablets for data entry.

Using the rectifier output example, a user could create a web-based database, with distinct user logins, and the ability to read, write, and edit entries on a per-user basis, in the course of an afternoon.

Maintainable

No survey sheet survives contact with an operating pipeline system. The addition or removal of individual test stations, or entire pipeline systems, are routine tasks in maintaining a CP data set.

When implemented in Excel, this usually takes the form of adding or deleting entire rows, which can break traceability between surveys, and makes historical data comparisons difficult.

Databases are built around the idea of managing changing data sets; individual measurement locations can be flagged as "inactive" or "historical" and omitted from further reporting.

The task of applying a common analysis to multiple Excel sheets is largely a copy and paste exercise, inserting an equation or graph into the files of interest. If an error is discovered, or an additional analysis is desired, the copy and paste exercise is repeated.

When a common analysis is applied to CP data in a database, the analysis is define in one place, and the user selects the data to which it is applied. Changes in the analysis happen in one place, and the copy and paste exercise is eliminated.

CASE STUDY:
REVIEW OF ER CORROSION PROBE DATA FROM TWO DIFFERENT SOURCES

To demonstrate the difference in usability of unstructured vs. structured CP data, a case study involving the review of two data sets from Electrical Resistance (ER) corrosion probes will be discussed. The goal of this review was to visualize corrosion rates for consideration in a larger pipeline integrity analysis (i.e. *"Just show me a graph."*).

Dataset 1 is a typical set consisting of manually measured readings, entered into a Word document and distributed via email, and Dataset 2 consists of an export from a typical remote monitoring website.

Dataset 1

Dataset 1 relates to with a set of test stations with soil side ER corrosion probes. Measurements were collected manually from these test stations to supplement CP structure-to-electrolyte potential data with a corrosion growth rate.

Dataset 1 was compiled in the following process:

1. Once a month, field crews would manually survey each of the six ER probes during other routine ROW inspections, and record this data on pen and paper.
2. The data would be transcribed into tables in a Word document, with a new file created each month.
3. These Word documents were stored in an unknown location by a field supervisor.
4. The Word documents were sent to the author as attachments to a series of emails.
5. The author transcribed the readings from each Word file into a database.

An example of a single month's report is shown in Figure 1, though the format changed several times through the reporting period.

KP	Special	Check
72	339	814
132	101	783
162	76	820
210	67	802
270	74	831
324	88	828
Calibrator	211	803

Figure 1: Typical Monthly Report for Dataset 1 (in Word Format)

It took roughly four hours to compile 18 years' worth of monthly data, and five minutes to do the review and analysis; this labour imbalance is not acceptable.

Implementation of the data governance described previously could have combined all of these five steps into one:

1. Once a month, field crews manually survey each of the six ER probes and enter the data directly into an online database that is accessible by whoever needs it.

[N.B. the author is not opposed to the use of pen and paper for *some* survey data, but rather seeks improvement on the storage and transmission methods of the data itself.]

Ultimately, the review of the data itself had identified data integrity concerns (e.g. identical data and or missing data, possible malfunctioning equipment), and meaningful analysis could not be performed.

Even though there was a formatted structure to how the data was visually presented (albeit in a Word table), fundamentally there were no data governance controls enforced on the data itself after collection. Under this method, there isn't an opportunity for a data structure to be deployed when raw data is collected without rigidity and uniformity, nor when the tabulated data is reported.

Dataset 2

Dataset 2 relates to a system of five remotely monitored ER corrosion probes installed within a pipeline pump station. By virtue of being measured by remote monitoring units and reported to a website, Dataset 2 was already uniform and accessible; the only way to alter the data format is by physically visiting and reconfiguring each of the remote monitoring units.

An example of the data presentation for a single probe is shown in Figure 2. The data itself is provided to the user by the improved method suggested for Dataset 1; it is fit for use literally at the click of a download button, as opposed to requiring hours of data consolidation. Note that the underlying data is available as a raw database table.

Date Time	Thickness (µm)
26-Aug-2018-22:01	97.60595103
26-Aug-2018-23:01	97.60278832
27-Aug-2018-00:01	97.60272756
27-Aug-2018-01:01	97.60602305
27-Aug-2018-02:01	97.60048399
27-Aug-2018-03:01	97.60319496
27-Aug-2018-04:01	97.60163758
27-Aug-2018-05:01	97.60127125
27-Aug-2018-06:01	97.60253244
27-Aug-2018-07:01	97.60352906
27-Aug-2018-08:01	97.61355465
27-Aug-2018-09:01	97.60253737
27-Aug-2018-10:01	97.60299769
27-Aug-2018-11:01	97.59816209
27-Aug-2018-12:01	97.60317726
27-Aug-2018-13:01	97.60220782
27-Aug-2018-14:01	97.60159689
27-Aug-2018-15:01	97.60157611
27-Aug-2018-16:01	97.6044012
27-Aug-2018-17:01	97.60476108

Figure 2: Sample from Dataset 2 (Online Database Format)

Note that this data very well could have been recorded and entered manually by a human, the important observation is that the remote monitoring system meets the data governance uniformity and accessibility requirements.

OPPORTUNITIES & PITFALLS

The "structured data" approach is increasingly applied to CP data. The migration is very labour-intensive, but the benefits present themselves almost immediately; through continued use, pitfalls also emerge.

Structure Data Benefits ("What Works")

The simple virtue of having access to all CP data for all assets for all time has greatly reduced the labour required for data lookups. Requests from other Pipeline Integrity staff for CP assessments (e.g. "*What do the CP levels look like at MP 123? Have they been good for the past three years?*") are easily managed, have helped keep other integrity projects on track, and reduced the workload for the CP data owners in general.

Having the structured dataset has also made running large analyses a routine matter. For example, a review of groundbed resistance trends is a matter of applying Ohm's law and graphing the result; further analysis (i.e. linear regression, identifying those with a positive slope) can automatically identify those that are in the process of failing. This can has help to create prioritized and defensible groundbed replacement budgets across large CP systems (e.g. *"I've got these three groundbeds that are actively failing, here are the graphs to prove it"* vs. *"this groundbed is old and might fail"*).

Simply put: strong data governance has increased the usefulness of CP data beyond meeting a compliance requirement. This was the desire, and through significant effort and discipline, it has been realized.

Structured Data Pitfalls ("What Doesn't Work")

The formalization of CP data into a structured data set presents many benefits, however it obscures a single, critical pitfall: CP data is only useful if it is valid. That is, all of the issues related to accurate surveying (e.g. total current interruption, equipment calibration, etc.) still exist, but the presentation of CP data in a highly uniform, structured manner runs the risk of it being interpreted as more useful or valid than it really is; this is especially true for non-CP users who have access to this data through the previously touted database accessibility benefits.

One example of this would be the ER corrosion probe readings presented in Dataset 2. They were uniform and accessible, however if the remote monitoring unit had been incorrectly configured, some or all of the data could actually be invalid, even though it might appear acceptable.

To this end, rules around the sharing of certain kinds of CP readings, or even individual readings, must be considered. The nuance behind, for example, a highly electronegative pipe-to-soil potential may be apparent to an experienced CP user, however it may not be apparent at all to someone who is running an analysis algorithm across a thousand measurement points over five years, and they may not even be looking at the individual readings themselves.

Before any reading is allowed to be made "public" (i.e. beyond CP experts), it must first be vetted, and this is the area where those CP experts can share their experience. Any reading or even an entire survey that is not immediately reliable must not be available for use by other users. For example, pipe-to-soil potentials on a pipeline with known uninterrupted influence from a distributed galvanic anode system has limited reliability and should likely be kept private; IR-free coupon-to-soil potentials for that same pipeline will likely be more reliable, and could be made "public."

The experienced CP user must be the gatekeeper for their CP data, and that includes both bringing it in to the database, and letting it out to other users.

CP DATA MANIFESTO

Summarizing these concepts, regardless of the actual implementation method, a few key rules emerge:

Rule 1: Keep all data in the same place all the time.
Rule 2: Make all data uniform.
  Subclause 1: If you want data in a certain format, define that format yourself, have someone take a look to make sure it makes sense, carve it in stone, and give it out willingly
Rule 3: Trust a reading yourself before you let anyone else use it.
Rule 4: Use Excel as a last resort; go learn how to use a database instead.

Adherence to this CP Data Manifesto should be considered the bare minimum requirement for good data governance. Any further improvements, whether discussed in this paper or not, to further improve of data governance. The first step, however, is to define *your* goal for implementing data governance.

CONCLUSIONS AND RECOMMENDATIONS

Conceptually, the integration of the various types of CP data and other pipeline integrity data is fairly trivial, but requires close attention to how well-structured data is deployed before integration itself. It's not until one has well-structured data that they should begin analysis, or integrating the various data.

The abilities afforded to us by Big Data technologies can only be unlocked to us by the application of good, consistent data governance. This paper has demonstrated trivial analyses that would have not been worth the effort to do on a large scale without the work already put in place to achieve well-structured, accessible data sets. Observing the rules set forth, and enforcing these good data governance practices, are the only sustainable ways to improve the usefulness of CP data.