



2021

Your energy is precious

Guide to
healthy
reverse
osmosis data
for machine
learning

Hello

We have prepared this guide on instrumentation and data collection to ensure the best data analysis for your plant. The suggestions suit Synauta's machine learning to save energy and chemicals while optimizing reverse osmosis (RO).

Following this guide means you can save energy or chemicals faster, saving OPEX sooner. For instance, a 300,000m³/day plant could be losing up to \$8,200 every day optimization is delayed.

Instrumentation

Instruments are key to managing preventative maintenance for a plant. They also serve to collect and share data that can be used for machine learning and other software applications to improve performance. The following are key instruments that will help RO be machine learning ready.

Flows

All flows must be provided, or must be able to be calculated from other flows, e.g. permeate and reject flows provided with feed flow calculated. An instrument error can be identified faster when there are more instruments in use. If flows are calculated using concentrations and an assumed recovery, an incorrect reading from an instrument can go undetected.

Pressures

Pressures must be provided for major flows. The most important are:

- feed pressure
- reject pressure or differential pressure
- permeate pressure.

Changing pressures are a key factor in identifying fouling and limiting flows. Missing instruments can open a system to undetected fouling and operating outside of designed limits, which can also be detrimental to plant operations in general.

Feed water

Collect accurate feed water data (temperature and conductivity). Changing feed water conditions play an important role in a system's pressures and fouling and therefore cannot be assumed based on occasional handheld readings. Feed water data also plays an important role in maximizing energy savings for any plant.

SDI or turbidity

Silt Density Index (SDI) readings or turbidity readings if relevant. If the plant experiences major fouling events regularly, then a regular SDI measurement (preferably an auto-SDI instrument) will be important. Lowering recoveries is sometimes necessary if undissolved solid concentrations rise too high as this limits the amount of particulate matter passing through the membranes.

General intake

Trains (racks) should not share sensors unless they are on general intake lines. Instruments located only on one train's pipework can lead to one train operating blind if another train is down for maintenance.

Handheld readings

Avoid handheld instrument readings. Handheld readings, especially with different operator use, may introduce inconsistencies.

Machine-generated results are ideal; they show Synauta the fine detail required for machine learning to provide the most accurate picture of the plant's operating state at any given time.

Instrument recording frequency

The data shown below has a number of issues.

- Inconsistent time intervals for different instruments. Ideally these should be the same.
- Inconsistent time interval for the same instrument. These should be consistent intervals.
- Duplicate timestamps where there only needs to be one recording.

For machine learning implications, in this example the timestamps still match across all instruments however, some record with a higher frequency than others. This means there may be more data in between the matching timestamps. It requires additional data sorting, which can increase processing time and introduce a source of error.

Eventtime	Descriptor	Value	Eventtime	Descriptor	Value
5/15/19 0:34	PC Example Plant RO Permeate Conductivity	88.326185	5/15/19 0:57	PC Example Plant RO Feed Inlet Conductivity	1294.056
5/15/19 0:34	PC Example Plant RO Permeate Conductivity	88.326185	5/15/19 0:57	PC Example Plant RO Feed Inlet Conductivity	1289.278
5/15/19 1:18	PC Example Plant RO Permeate Conductivity	89.371341	5/15/19 2:37	PC Example Plant RO Feed Inlet Conductivity	1295.299
5/15/19 1:18	PC Example Plant RO Permeate Conductivity	89.371341	5/15/19 2:37	PC Example Plant RO Feed Inlet Conductivity	1296.211
5/15/19 2:12	PC Example Plant RO Permeate Conductivity	88.829729	5/15/19 3:12	PC Example Plant RO Feed Inlet Conductivity	1295.983
5/15/19 2:12	PC Example Plant RO Permeate Conductivity	88.829729	5/15/19 3:12	PC Example Plant RO Feed Inlet Conductivity	1295.146
5/15/19 4.23	PC Example Plant RO Permeate Conductivity	89.292929	5/15/19 5.43	PC Example Plant RO Feed Inlet Conductivity	1294.819
5/15/19 4.23	PC Example Plant RO Permeate Conductivity	89.292929	5/15/19 5.43	PC Example Plant RO Feed Inlet Conductivity	1297.422

File and data format

Data is the essence of machine learning and Synauta's system loves data in a CSV file format!

Many of the formatting challenges outlined in this section can be avoided if your data is in a CSV. If you are using Excel, we have included recommendations on what to avoid before exporting to a CSV.

Getting plant data into a CSV

Ideal: Your SCADA automatically downloads a CSV. The SCADA manufacturer can help automate this.

Option: If your data set has less than one million rows (the maximum number Excel allows) it can be downloaded to Excel and manipulated in a simple manner, then Excel can export the data to a CSV format. A macro can also be written within Excel to help you sort your data.

Synauta also works with customers who have direct, secure data transfers from SCADA systems.

Advantage of Comma-Separated Values (CSV)

Ideal: Comma-Separated Values file (CSV) with one consistent delimiter and in a single file. Note the CSV acronym can also stand for Character-Separated Values or Comma-Delimited files.

Why: You can export complex data from one application to a CSV file and import the data in that CSV to another application.

A CSV prevents more than one dataset in one column or the risk of introducing complex layouts (like Excel files can create). For machine learning, CSVs have a small file size and simple file layouts to make reading and processing data faster.

This example data set has inconsistent rows. In an XLSX file (before 2007 called XLS file), the data set may have complex formatting.

Excel formatting introduces limitless data layouts, which takes longer for software to be adapted to automatically read a specific Excel file.

Avoid

	A	B	C	D	E	F	G
4/5/20 12:00:05 AM							35
4/5/20 12:01:05 AM	12.4	9.2	3.1	9.4	8.4	6.3	15
4/5/20 12:02:05 AM							36
4/5/20 12:03:05 AM	12.5	9.6	2.8	10	9.3	6.6	18
4/5/20 12:04:05 AM							36
4/5/20 12:05:05 AM	13.1	10.8	2.9	9.8	9.5	7.2	17
4/5/20 12:06:05 AM							35
4/5/20 12:07:05 AM	13.3	10.8	2.7	9.5	8.6	7.1	18
4/5/20 12:08:05 AM							42
4/5/20 12:09:05 AM	13.1	10.4	2.8	9.2	9.5	7.2	17


One file, one worksheet tab

Ideal: Data saved to one file and in one worksheet tab (also known as a sheet).

Why: Splitting data across multiple files takes up unnecessary space for the repetition of dates. It also adds to processing time due to the file stitching required and potential to introduce errors due to this process.

For machine learning, multiple worksheet tabs are as clunky as multiple, separate files.

Ideal

 RO Data Oct-20 - Feb-21.csv

		TIT1101	FIT1102	PIT1101
1				
2	10/5/20 12:00:05 AM	25.0	983.2	75.89
3	10/5/20 12:01:05 AM	25.0	985.2	75.78
4	10/5/20 12:02:05 AM	25.0	985.3	75.73
5	10/5/20 12:03:05 AM	26.0	982.3	75.28
6	10/5/20 12:04:05 AM	25.9	982.7	75.26

Navigation: Oct-20 - Feb-21 +



Avoid

-  RO Data Dec-20.csv
-  RO Data Feb-21.csv
-  RO Data Jan-21.csv
-  RO Data Nov-20.csv
-  RO Data Oct-20.csv

	A	B	C	D	E
1		TIT1101	FIT1102	PIT1101	
2	10/5/20 12:00:05 AM	25.0	983.2	75.89	
3	10/5/20 12:01:05 AM	25.0	985.2	75.78	
4	10/5/20 12:02:05 AM	25.0	985.3	75.73	
5	10/5/20 12:03:05 AM	26.0	982.3	75.28	
6	10/5/20 12:04:05 AM	25.9	982.7	75.26	

Navigation: Oct-20 | Nov-20 | Dec-20 | Jan-21 | Feb-21

Single date/time column

Ideal: Column 1 in any file should be a single date/time column. All other columns should be plant instrument readings.

Why: When a single date/time column exists, data processing is much faster. When multiple date/time columns exist for every instrument it increases the file storage size, making data processing slower. A single date/time column also enables a higher frequency of set points to be recommended, creating greater savings for your reverse osmosis process.

Ideal

	TIT1101	FIT1102	PIT1101
4/5/20 12:00:05 AM	25.0	983.2	75.89
4/5/20 12:01:05 AM	25.0	985.2	75.78
4/5/20 12:02:05 AM	25.0	985.3	75.73
4/5/20 12:03:05 AM	26.0	982.3	75.28
4/5/20 12:04:05 AM	25.9	982.7	75.26
4/5/20 12:05:05 AM	25.1	983.1	75.01
4/5/20 12:06:05 AM	26.0	985.6	75.09
4/5/20 12:07:05 AM	26.0	985.9	75.89
4/5/20 12:08:05 AM	25.5	985.1	75.23
4/5/20 12:09:05 AM	25.7	985.6	75.67
4/5/20 12:10:05 AM	25.7	984.2	75.76

Avoid

TIT1101 Time	TIT1101	FIT1102 Time	FIT1102	PIT1101 Time	PIT1101
4/5/20 12:00:05 AM	25.0	4/5/20 12:00:05 AM	983.2	4/5/20 12:00:05 AM	75.89
4/5/20 12:01:05 AM	25.0	4/5/20 12:01:05 AM	985.2	4/5/20 12:01:05 AM	75.78
4/5/20 12:02:05 AM	25.0	4/5/20 12:02:05 AM	985.3	4/5/20 12:02:05 AM	75.73
4/5/20 12:03:05 AM	26.0	4/5/20 12:03:05 AM	982.3	4/5/20 12:03:05 AM	75.28
4/5/20 12:04:05 AM	25.9	4/5/20 12:04:05 AM	982.7	4/5/20 12:04:05 AM	75.26
4/5/20 12:05:05 AM	25.1	4/5/20 12:05:05 AM	983.1	4/5/20 12:05:05 AM	75.01
4/5/20 12:06:05 AM	26.0	4/5/20 12:06:05 AM	985.6	4/5/20 12:06:05 AM	75.09
4/5/20 12:07:05 AM	26.0	4/5/20 12:07:05 AM	985.9	4/5/20 12:07:05 AM	75.89
4/5/20 12:08:05 AM	25.5	4/5/20 12:08:05 AM	985.1	4/5/20 12:08:05 AM	75.23
4/5/20 12:09:05 AM	25.7	4/5/20 12:09:05 AM	985.6	4/5/20 12:09:05 AM	75.67
4/5/20 12:10:05 AM	25.7	4/5/20 12:10:05 AM	984.2	4/5/20 12:10:05 AM	75.76

Chronological timestamps

Ideal: Instrument data in chronological order.

Why: Timestamps must be consistently timed apart with a max interval of 5 minutes (ideally 1 data point per minute). Having more data than once per minute is okay, however it greatly increases the file size for little return. You will end up with file sizes in hundreds of gigabytes (GB) for just one year, whereas 20-30GB is ample. This will reduce your data storage costs.

Ideal: Timestamps must line up across instruments.

Why: Gives an instantaneous global image of the plant's current operation.

Timestamp day/month/year order

Ideal: A consistent order e.g. day, month and year could be presented as 05/11/2020, 11/05/2020, 2020/11/05, 2020/05/11 as long as the order is consistent throughout the data.

Why: Changing the order will introduce errors and increase processing time.

Avoid negative values

Ideal: Values to be a 0 when a plant is off, instead of a negative value like -2. Generally, a negative instrument value means the instrument needs rescaling.

Why: 0 is easier to identify when the plant is off and has no flow / pressure.

Column headings

Ideal: Each column should have a descriptive heading, either the instrument tag number or a descriptive name that can be used to easily identify the instrument. The best approach is when a plant's column headings match the P&IDs. This makes data analysis extremely fast.

Why: Speeds up data processing time so you can receive analysis and optimized recommendations faster.

Language

Ideal: Column headings with instrument numbers from the P&ID are the best option. Otherwise, file names in consistent language e.g. all English, all Spanish, etc.

Why: Consistency avoids potential errors and accelerates analysis.

Characters

Ideal: Standard English characters e.g. AM instead of $\mu\mu$.

Why: Makes it easier for comprehension and processing the way a machine learning data pipeline is arranged. Synauta would be required to make additional adjustments to accommodate such symbols, which increases processing time.

	Ideal			Avoid			
Timestamp	TIT1101	FIT1102	PIT1011	Timestamp	TIT1101	FIT1102	PIT1011
4/5/20 12:00:05 AM	25.0	983.2	75.89	4/5/20 12:00:05 $\mu\mu$	25.0	983.2	75.89
4/5/20 12:01:05 AM	25.0	985.2	75.78	4/5/20 12:01:05 $\mu\mu$	25.0	985.2	75.78
4/5/20 12:02:05 AM	25.0	985.3	75.73	4/5/20 12:02:05 $\mu\mu$	25.0	985.3	75.73
4/5/20 12:03:05 AM	26.0	982.3	75.28	4/5/20 12:03:05 $\mu\mu$	26.0	982.3	75.28
4/5/20 12:04:05 AM	25.9	982.7	75.26	4/5/20 12:04:05 $\mu\mu$	25.9	982.7	75.26

Units

Ideal: If your file does not use instrument numbers from the P&ID for row headings, include units in the heading of each column.

Why: Helps quick reference of data to avoid errors and saves time when confirming units. Also ensures consistency across all instruments. You can check unit settings by looking at the sensor's range and how it is set to record in SCADA.

Comments

Ideal: Only instrument values should be in each cell, no comments or units.

Why: Comments will invalidate the value in the cell and data-reading software cannot understand these comments.

Final tips

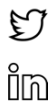
It's always best to work with the data assets you have now and enable subsequent projects to inform data collection efforts.

When you start working with Synauta we create a Machine Learning Ready Report for your plant. This includes analysis of the plant instrumentation effectiveness, data health and membrane performance, with assistance for operators and plant managers on data rectification to maximize OPEX savings using machine learning.

Free Machine Learning Ready CSV template

Contact us for a free CSV template in an ideal, machine learning ready format. We're also here to answer questions and welcome any feedback on this guide.

info@synauta.com
+1 403 861 2036
synauta.com



The information and data contained herein are deemed to be accurate and reliable and are offered in good faith, but without guarantee of performance. Synauta assumes no liability for results obtained or damages incurred through the application of the information contained herein. Customer is responsible for determining whether the products and information presented herein are appropriate for the customer's use and for ensuring that customer's workplace and disposal practices are in compliance with applicable laws and other governmental enactments. Specifications subject to change without notice. Copyright © 2021 Synauta. All rights reserved. Version 202104.