

Improving geoscience data access with automated workflows

Paritosh Bhatnagar, Chris Hanton and Phil Chalmers
Ikon Science

Summary

The diversity of subsurface data generated, interpreted, and utilized throughout an asset's lifecycle has dramatically increased in recent years, as new acquisition and monitoring technologies promote a digital oilfield. At the same time the number of dedicated data managers has decreased across our industry as a series of commodity price downturns have resulted in significant headcount reductions. The conflicting result of more data but less resources to effectively manage it poses severe challenges in accessibility, analysis, and usability of information to generate insights. To address these issues, this paper introduces automated data management workflows which introduce scalability and speed to key subsurface data management processes without the need for increased headcount by operators. Data dragged into formalized staging area triggers a number of configurable and automated processes that cover data categorization, parsing, assessment, enrichment and loading of both structured and unstructured formats for a variety of complex data types. Implementation with two Canadian operators show how automated workflow can enhance storage and access to definitive versions of well data, fostering a robust corporate data management strategy.

Introduction

Throughout an asset's lifespan, data takes on numerous forms and types as it is gathered, interpreted, and employed. This valuable data is ultimately scattered across various platforms, as data is fragmented and stored across different applications, systems and folder locations making it difficult to access, analyze, and utilize effectively. This leads to low data awareness within an organization. As a result, users consistently ask the following questions:

- “What data is available for a given well?”
- “What is the quality of my data?”
- “Where is my data stored?”
- “Is this the correct version of my record?”
- “How to get the data in the right format?”

Users spend considerable time searching for data and trying to determine its origin, quality, and its trustworthiness. Consequently, low quality data can lead to inaccuracy and poor decision-making.

There have been efforts to improve the siloed nature of data storage and provide a single data platform which many subsurface applications could potentially connect to (Kaur et al., 2021, Tomlinson et. al 2022). However, they lack storage of complex data types (such as core data) or ability to search for data and assess its quality. If a platform does handle different data types, it lacks integration with other databases and applications ultimately causing bottleneck for data access.

In this paper, we showcase the autoloaders – an automated data ingestion technology, designed to streamline geoscience data population and improve accessibility. The tool categorizes and classifies information based on contents and metadata properties and applies business rules, quality checks and standardization rules through automated processes. Two case studies are presented showcasing autoloader configuration, expandable data model support and post-load analysis.

Method

As noted, autoloader is an ingestion tool that handles complex subsurface data and provides high quality of trust and confidence for geoscience data access. It is a type of data crawler that systematically searches within network drives to discover and extract content, which can then be searched or analysed. It does so by first extracting the well name or UWI based on folder name, file name or its content. If the well does not exist within the database, it can auto-create the well based on its metadata extraction. As part of the automation, it also extracts other well metadata properties from various file formats. These include, but not limited to XLS, CSV and LAS files. Given subsurface data comes in various types and formats, it is especially designed to intelligently parse out data based on custom business rules and run data quality checks that can range from curve/unit aliasing, flagging curves with unexpected data values, null value analysis, R2 checks to metadata validation. This ensures data ingested into the system is of high quality and meets pre-defined standards. It also honours a check for data duplication using a MD5 checksum based on file name or its content.

As it scans for files on a scheduled basis, it categorizes data based on its content or pre-defined templates. For example, a specific data model can be configured if certain sets of curves are present, categorizing the dataset as triple combo vs quad combo. Similarly, it can reference custom spreadsheet templates and automate the standardization of core data along with capturing its source, versioning, and data type. This ensures a consistent data model and gives the users the ability to search and access the data that is most relevant to them. If the data does not meet the standards, it is sent to the quarantine folder for further review. Figure 1 shows a schematic of an XLS file type and the logic behind data extraction as the file is processed through the autoloader. The basic steps are:

1. Detect for duplication and parse data from the source object (e.g., file, connected database, integrated application)
2. Run against various templates and categorize the data
3. Apply business rules and QA/QC checks
4. Load data or reject to quarantine area depending on if rules are passed/failed
5. Notify load outcome via reports/email

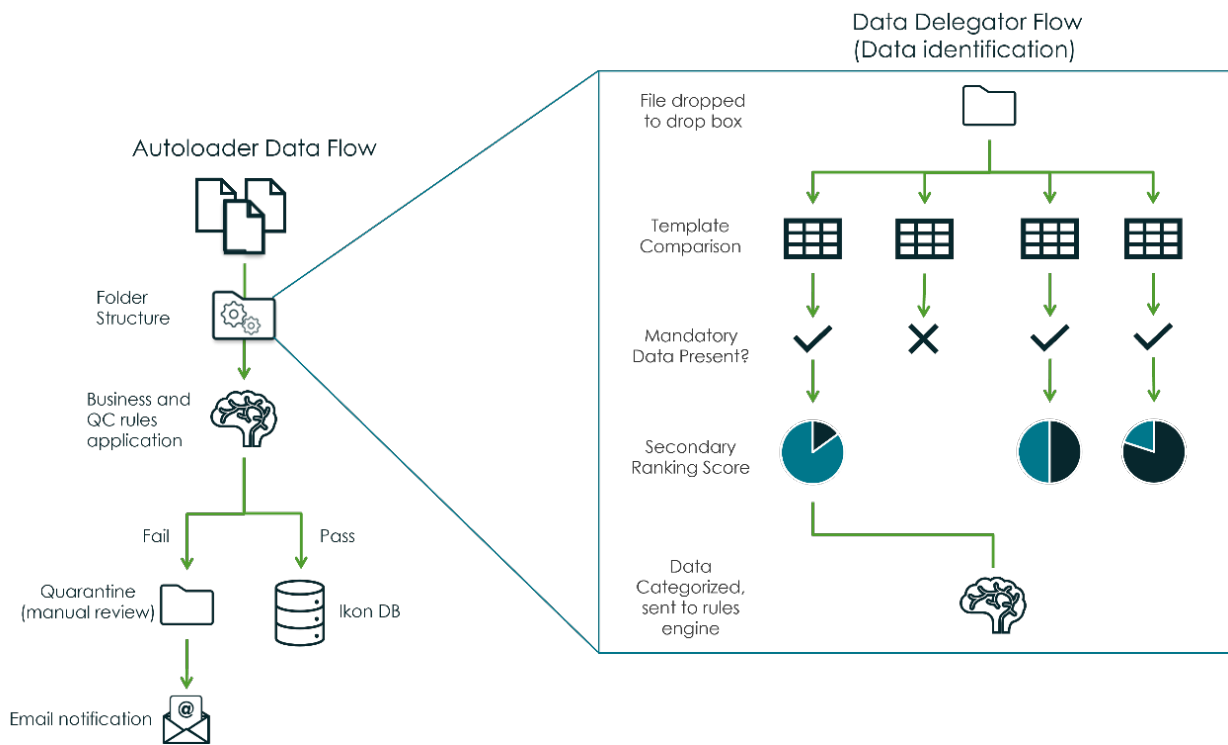


Figure 1: Autoloader data flow for an example XLS file type following business QC rules

The time spent to extract data can take anywhere from seconds to couple of minutes, depending on the content and size of the file. For example, autoloader can process and extract data from 85 files ranging from XLS, LAS, PDFs and Images (plugs and SEM photos) within 2 minutes.

The above workflow validates that the data meets QA/QC and business rules upon loading, ensuring that data accessed through the system is of a consistently high standard. Once the qualified data is loaded into the database, structured data in the form of table views can be exposed to allow analytics applications such as Spotfire or Power BI to utilize the data and perform in depth data analysis and generate reports. This clean set of data can further be utilized for any ML/AI workflows.

Case Study 1: AutoQC Workflow for Log Promotion

In this implementation, autoloader was deployed for a mid-size Canadian operator to standardize raw and interpreted data, understand log data quality, and promote curves based on custom business rules configuration. As data was auto ingested, autoloader went through the following rules:

- Unit Check/Conversion: Use Unit alias tables to apply unit conversions for the different log types
- Curve Range: Check for data values and ensure they are within the log limits specified in the curve alias table (system default where curve mnemonics are captured); further check for spikes and null values
- Data Depth Ranges: Scan the various vertical ranges of the logs and for completeness (longest continuous log); If bottom of log is available, take the log that has largest depth range
- R2 check: if two similar log types exist in a dataset, compute an R2 check and evaluate how similar the logs are
- Promotion of multiple versions of log: if the above checks are passed, promote curve and calculate its quality score

Following the checks outlined above, once the files are put in the staging area the autoloader identifies the curve and looks up the appropriate units for that curve type. If units are correct and the log values are within the log range, it promotes it and loads it into the database. If units are incorrect, autoloaders convert it using a reference unit alias table. If the curve values are out of range, it checks the percentage of samples that are out of range. For a certain threshold, it either rejects it and moves it to the quarantine folder or loads the data. If similar log types are present in the file (for example 3 versions of PHIE log), it checks for overlaps between similar curve and by how much the overlap is. For a certain threshold value and an R2 check in place, it promotes or rejects the curve for further review. As the logs are promoted and loaded in the database, autoloader calculates a final QC score to expose data quality.

This custom configuration allows promotion of multiple versions of logs from raw data, exposes curve quality and standardizes various log-based data on the fly. Moving forward, autoloader can apply the above checks on any given LAS dataset and ensure high quality data is auto ingested for users to easily access and use them in their workflows.

Case Study 2: Subsurface Well and Documents Records Management

For this implementation, the business unit was interested in a comprehensive master well datastore for various data analysis and interpretations (generated from subsurface applications and tools), provide means to load documents and categorize them, manage file contents and capture relevant metadata information for easy querying. The dataset included 28,500 wells (well headers and deviation surveys) imported from Geoscout, 4500 datasets (ranging from logs, core data, formation tops, thin sections, raster images, core description and photos). The aim was to automate the data ingestion process and house this data in a single environment that enables both collation and cleanse of data prior to utilization in geoscience workflows.

The autoloaders were configured for loading core data, where custom templates were defined to handle data coming from various vendors. Similar to previous study, curve aliases were referenced during the data ingestion process. Spreadsheet templates ensured data

categorization such as data model definition, curve standardization and metadata extraction were captured. As part of the process, further checks were implemented such as mandatory curves present in the file, running templates on each excel tab and merging similar datasets. If any of the business rules were broken, the file was sent to the quarantine area. Examples of some templates include routine core analysis, core gamma, x-ray diffraction and mercury injection capillary pressure from Weatherford, Corelab and Agat labs.

Another powerful feature of autoloader is its ability to archive files and use direct links to where the files are stored on the network drive for easy access and retrieval. To manage unstructured data for the business unit, autoloaders processed and archived various documents based on where they were placed in the staging area. In this configuration, folders were defined based on well UWI followed by document type (core description or drilling report). Additionally, all the original files (LAS, XLS) were automatically attached at the dataset and well level for easy retrieval.

Once all the data is loaded, geoscientist can now perform multi-scale visualization of various data types to enhance their interpretation in an intuitive environment, that may have been challenging to do so in the past (Figure 4).

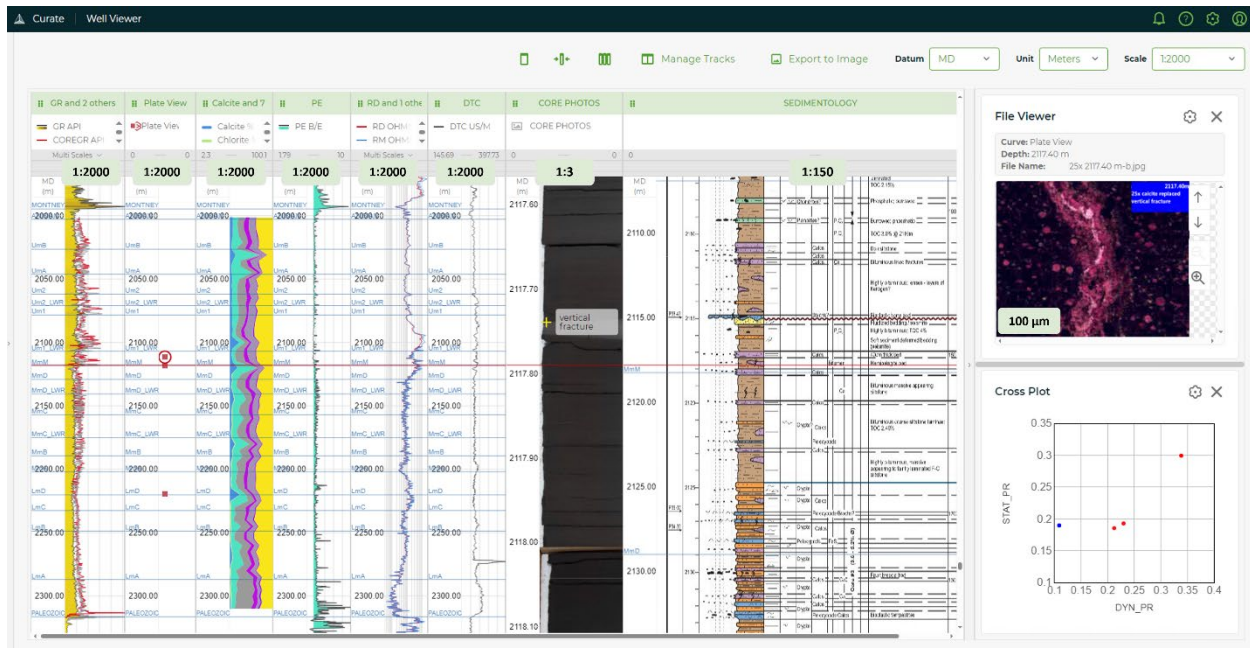


Figure 4: Multi-scale visualization of various data types auto-ingested as part of the autoloaders

Conclusions

Automated data loading workflows increase speed and consistency of data population and improve everyday productivity. The time spent ingesting data is greatly reduced (from months to few days) providing not only efficiency in data loading but retrieval of information and instilling confidence in the quality and accuracy of the data. Autoloader's customizable business rule addresses data quality, data duplication, and governance aspects, ensuring a consistently high standard of data to be loaded into the system. Automatic QC score calculation ensures data meets predefined standards while providing transparency in process building to gain data confidence. With clean and consistent data types, these datasets can effectively be utilized for ML/AI workflows.

Acknowledgements

Thanks to Phil Chalmers for designing the autoloader and our colleagues at Ikon Science for their technical input.

References

Kaur, A., P. Hodson, S. Vallabhaneni, E. Wild, and M. R. Satapathy, 2021, And the walls came tumbling down: OSDU is dismantling the data silos of energy companies once and for all: Accenture, https://www.accenture.com/_acnmedia/PDF-163/Accenture-AndThe-Walls-Came-Tumbling-Down.pdf, accessed 26 July 2022.

Tomlinson, J., P. Bhatnagar, and C. Hanton, 2022. A roadmap to accelerate OSDU adoption. *The Leading Edge*, 41(9), pp.647-651. doi: <https://doi.org/10.1190/tle41090647.1>