D5.2 Data Lifecycle and Curation Strategies

Deliverable No	D5.1
Work package No. and Title	WP5 OPTIDRILL Well, Drilling, stimulation and completion database
Version - Status	V1.0 – draft
Date of Issue	25/10/2021
Dissemination Level	PUBLIC
Filename	OptiDrill-D5.1-Data Lifecycle and Curation Strategy

Disclaimer: Any information on this deliverable solely reflects the author's view and neither the European Union nor CINEA are responsible for any use that may be made of the information contained herein.

This project has received funding from the European Union's Horizon 2020 research and innovation action under grant agreement No 101006964

DOCUMENT INFO

Authors

Author	Organization	e-mail
Towseef Ahmed	Technovative Solutions Ltd	towseef@technovativesolutions.co.uk
Mohammad Ashadul Hoque	Technovative Solutions Ltd	ashadul@technovativesolutiuons.co.uk

Document History

Date	Version	Editor	Change	Status
07/10/2021	v1.0	Mohammad Ashadul Hoque	Initial draft	Draft
14/10/2021	v1.0	Mohammad Ashadul Hoque	Final draft	Draft
25/10/2021	V1.0	Shahin Jamali	Submission ready	

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The OPTIDRILL project aim to develop a drilling advisory system utilizing novel sensor and machine learning methods to predict ROP, lithology, drilling problems, well completion and enhancement, and unite those methods under one system to enable drilling process optimization and intelligent decision making. The project's main objectives therefore include:

- Develop enhanced drill monitoring systems based on measurement while drilling (MWD) systems and acoustic- and vibration-based sensors.
- Develop automated machine learning-based analysis methods to predict drilling parameters using sensor-based data-driven models.
- Develop a real-time drilling monitoring and optimization tool as a unified system to combine the existing data with the newly developed methods.
- Develop coupled drilling optimization models to reduce geothermal drilling costs
- Develop sustainability model of OPTIDRILL.

The current deliverable describes the consortium's strategy for data lifecycle and curation strategy of the data collected, processed and/or generated because of the above-mentioned activities. The report includes information on 1) the strategy of handling of data during and after the end of project 2) methodology and standards applied and 3) data sharing, preservation, and security. The OPTIDRILL data Lifecyle and curation strategy is a *living* document ensuring a dynamic data management lifecycle and integration of updates over the course of the project for any significant changes to the consortium's policies and/or the addition of new datasets. The structure of this document is as follows: Section 1, presents theoretical introduction to the Data Lifecycle and curation as well as the description of the Lifecycle and curation methodology to be used within the OPTIDRILL project. Section 2 describes how each phase of the methodology is applied in the project and, finally, Chapter 3 will present conclusions to the Data Lifecycle and Curation Strategy.

1 Data Lifecycle and Curation

Research projects require the incorporation of data management principles and best practices for the proper execution. Data lifecycle illustrates stages of data management and describes the flow of data from start to finish. Data Lifecycle Model (DLM) offer a high-level framework of actions and processes that must be undertaken at different stages. The goal is to optimize data management, from efficient organization to elimination of any kind of waste, to provide a meaningful, high-quality data according to the user's expectation and requirements. The DLM seeks to assist scientists in anticipating and planning for specific actions that need to be taken to manage the data. Some models have been proposed to manage data for a specific scientific field or project and just consider some elements of a complete lifecycle of data. Other models focus on specific data phases, such as data curation or data preservation. A DLM represents the sequence of stages that a particular unit of data goes through from its collection to its eventual archival and/or deletion, at the end of its useful life. Although, the different data management models differ among themselves, most of the models have several steps in common; facilitating several steps that allow to manage, document, protect, and share data in an understandable manner.

Data curation is the process of collecting, organising, annotating, labelling, describing, cleaning, preserving, and maintaining data for use. It is the end-to-end process of creating good data through the identification, formation and management data through its lifecycle, from creation or collection and initial storage to that time when its archive for future research or disposed. The goal of data curation is to ensure that data is 'cared for' throughout its lifecycle so that its findable, accessible, interoperable, and reusable to get maximum possible value. Data curation is concerned more about maintaining and managing the metadata itself rather than the actual data. Effective data curation is especially important in ensuring ML/AI training data is prepared for processing i.e. data is machine readable, reliable and unbiased. Data curations is the place where experts can add their knowledge to the training data for effective ML/AI models. Good data curation practices are therefore essential for ensuring that research data are of high quality, findable, accessible and have high validity.

In this deliverable, we present a data lifecycle and curation model (DLCM) that has commonalities with other frameworks such as DataONE Data Lifecycle[1], Digital Curation Centre model[2], Information Lifecycle Management (ILM)[3], USGS data lifecycle model[4], but is aimed at OPTIDRILL project specifically. The DLCM promotes a lifecycle approach to the management of digital materials, to enable their successful curation and preservation from initial conceptualisation to either disposal, or selection for reuse and long-term preservation. A lifecycle approach ensures that all the required stages are identified and planned, and necessary actions implemented, in the correct sequence. This can ensure the maintenance of authenticity, reliability, integrity and usability of digital material, which in turn ensures maximisation of the investment in their creation.

[1] https://www.dataone.org/data-life-cycle

[2] http://www.ijdc.net/article/view/69/48

[3] https://www.gartner.com/en/information-technology/glossary/information-life-cycle-management-ilm

[4] https://pubs.er.usgs.gov/publication/ofr20131265

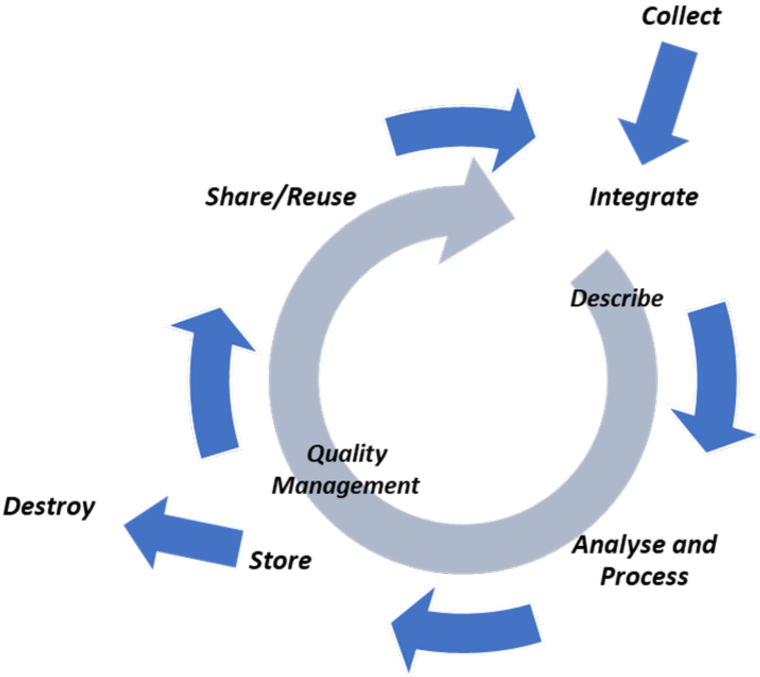

***Figure 1**: OPTIDRILL Data Lifecycle and Curation Model (DLCM)*

The following elements of the DLCM are highlighted from the figure above:

1) **Collect** represents activities where new or existing data is collected, generated, considered and evaluated for use. Researchers should define data acquisition techniques to address research questions. The outputs of this phase are the project's data inputs.
2) **Integrate** represents activities where newly collected datasets are incorporated into a data management system. This may involve initial processing so that they share formats and granularity, or relevant fields are correctly mapped. The researcher also needs to ensure integration at dependency level e.g. controlled vocabularies used in the data.
3) **Analyse and Process** represents ETL (extract, transform and load) activities of the dataset and calibration to prepare for analysis, exploration and interpretation of data including summarisation, statistical analysis to produce relevant results and information. Data management during this stage improves the efficiency of the activities, preserves metadata, documentation and creates the foundation for future research.
4) **Store** represents activities long-term preservation of data, metadata, additional results, storage formats, metadata, ontology and any additional documentation, to ensure availability and reuse.
5) **Share/Reuse** represents sharing data and metadata to designated users and re-users on a day-to-day basis. This may be in the form of publicly available information or with select users through robust access control and authentication procedure.
6) **Destroy:** Data remediation features will be supported after data encryption, shredding, or quarantining. This stage also will be linked to the generation of backups, prevent deletion of expired back-up data unless this is within the policy of data management.

The inner circle represents DLCM processes as defined below:

- **Describe** emphasizes the need to maintain proper metadata and documentation throughout the data lifecycle. This ensures the possibility of evaluating, validating, and replicating the research work from the data provided.
- **Quality Management** seeks to ensure quality of data throughout its lifecycle. In this way it is possible to verify, at an early stage, possible errors that affect the final result of the investigation.

2 OPTIDRILL Data Lifecycle and Curation Strategy

In OPTIDRILL the Data Management Plan (DMP) was defined in Deliverable 5.2 where the consortium's strategy for management of the data collected, processed and/or generated was described. The report included information on:

- Handling of data during the project
- Types of data collected, processed and/or generated
- Methodology and standards applied
- Data sharing, archiving and security

This document describes how data will be managed throughout the lifecycle. As part of making research data findable, accessible, interoperable and re-usable (FAIR), this should include the following information:

- Description of data to be collected, processed and/or generated
- Description of methodology and standards which will be applied
- Data Sharing information
- Description of curation and preservation of the data

2.1 Collect

Datasets collected and generated during OPTIDRILL project are of four categories; a) drilling datasets from consortium partners which are commercially sensitive in nature, b) public datasets such as "Netherlands and Dutch continental shelf", c) simulated drilling data generated during the project and d) What-if scenario data generated during the project. Fraunhofer has dedicated a directory in their ownCloud installation for OPTIDRILL project. Commercially sensitive drilling dataset will be encrypted by corresponding consortium partner and stored in their private directory integration into OPTDRILL data management system.

2.2 Integrate

First step in utilising collected datasets are their integration in a data management system. We have chosen CKAN[5] as our data management system. Due to commercially sensitive nature of some of the datasets the CKAN is installed in an air-gapped PC which means that collected data must be transferred by hand into this system before any integration can take place, including decrypting encrypted datasets.

Through the planning stages of the OPTIDRILL project it was determined that a curation of drilling related content was essential to enhancing the overall data quality and assurances throughout project. The most useful data has metadata about its creation, content and context. When metadata is well structured, uses consistent names and agreed upon vocabularies, it enables machine readability, aggregation, integration and tracking across datasets: allowing for Findability, Interoperability and Reusability. Hence, extensive metadata describe each dataset will be used in the integration phase.

2.3 Analyse & Process

Along with the increasing variety and heterogeneity of data sources, getting the data ready for analysis is critical. This may involve initial processing of datasets so that they share format and granularity, or so that relevant fields are mapped correctly. Datasets must be catalogued and connected before they can be used for analysis. Duplicate data and blank fields need to be eliminated, misspellings fixed, columns split or reshaped, and data need to be enriched with data from additional or third-party sources to provide more context. Cleaning data is one of the most important tasks under data curation. Data analysis involves various activities associated with exploring and interpreting data which include statistical Analysis, visualization, interpretation etc.

[5] CKAN is an open-source DMS (data management system) for powering data hubs and data portals. CKAN makes it easy to publish, share and use data. It powers hundreds of data portals worldwide.

Processing data involves various activities associated with the preparation of collected datasets including validation, transformation for generate derived datasets. These activities are to ensure best quality of the data used or created within the OPTIDRILL project. One of the activities that will performed is conversion data from one format to another suitable for ingestion by machine learning processes. Transformed data will be loaded into the dataset as a different resource to keep track of source and derived data.

Recording of how data is process and analysed will be kept for reproducibility and assessment of research quality.

2.4 Store

Store involves actions and procedures used to ensure long-term sustainability and accessibility of data and repositories. Encrypted drilling datasets from consortium partners and what-if scenario data will be stored in Fraunhofer ownCloud. Unencrypted drilling datasets from consortium partners, public datasets, simulated drilling, and intermediary dataset generated throughout the project will be included in the local CKAN installation in an air-gapped PC. CKAN facilitates users to set metadata, file descriptions, link together relevant data/datasets, i.e. organise and group data together with ease of access and availability. The use of custom APIs in CKAN allows the user a novel way to sort out and organise the entire database using the data/metadata.

2.4.1 Preservation

CKAN data management system ensures that data remains authentic, reliable, and usable while maintaining its integrity through metadata describing the data, source of the data and preventing user modification and deletion.

2.4.2 Backup & Security

Backup and security involve protecting data from accidental loss, corruption, and unauthorized access. CKAN authentication mechanism will be used to prevent unauthorised access. Air-gapped nature of CKAN installation also helps in securing the data; one needs physical access in addition to CKAN credentials to access data. Fraunhofer backup and recovery policy will be applied for the CKAN installation.

2.5 Share/Reuse

Most of the collected data are either confidential hence cannot be shared with external users or they are already publicly available from their respective authorities. Intermediary data generated during the project through ETL processed are very specific to OPTIDRILL development and will not be any use to external parties. Hence, they will not be shared. What-if scenario data generated during the project will be shared with external users for research purpose and give industry confidence in OPTIDRILL drilling advisory system.

Encrypted drilling data from consortium members stored at Fraunhofer ownCloud private directories will be used for populating CKAN data management system on the air-gapped PC only. Due to the nature of air-gapped PC, all the unencrypted drilling data from consortium partners, collected public datasets and simulated drilling data and corresponding metadata are accessible only to designated IEG user for the development of OPTIDRILL technologies.

2.6 Destroy

Datasets collected and generated during OPTIDRILL project are of four categories; a) drilling datasets from consortium partners which are commercially sensitive in nature, b) public datasets such as “Netherlands and Dutch continental shelf”, c) simulated drilling data generated during the project and d) What-if scenario data generated during the project. As drilling dataset from consortium partners are for OPTIDRILL use only, they will be destroyed at the end of the project. Public datasets are already shared by their respective authority and simulated drilling data are very specific to OPTIDRILL and these will also be destroyed at the end of the project. All backups will also be destroyed at this stage.

2.7 Cross-Cutting Methods

Cross-cutting elements describe activities that must be performed continuously across all stages of data lifecycle.

2.7.1 Describe (Metadata, Documentation)

Throughout the data lifecycle, metadata and documentation must be created and updated to reflect actions taken upon the data. Metadata to describe the data, the source etc will be added to all data collected and generated within the project. This enables the data to be managed and preserved for future references (within the duration of the project). The examples of metadata that could be used is described as follows. Keeping either the above table or bullet point list below

- Title: Title of the data set (also able to edit a specific URL as well)
- Description: User is able add a description about the dataset (can apply Markdown Formatting if required)
- Tags: Identifiers to quickly locate the data set by searching for key words/values
- Licence: Licence option can be changed as defined by opendefinition.org
- Organisation: Name of the organisation affiliated to the dataset
- Visibility: Private or Public data
- Source: Source of the dataset: literature/webpage/etc
- Version: Describes if the dataset has been edited/changed compared to the initial data
- Author: Person/persons who created the dataset
- Maintainer: Person/persons who will maintain the dataset throughout the lifecycle

With the use of CKAN data management system the metadata can easily be applied to datasets and will allow users to access and use the required data efficiently.

2.7.2 Manage Quality

Data-quality management is a process where protocols and methods are employed to ensure that data are properly collected, handled, processed, used, and maintained at all stages of the lifecycle. The OPTIDRILL project aim to develop a drilling advisory system utilizing machine learning methods to predict ROP, lithology, drilling problems, well completion and enhancement, which has critical dependency on high quality data, hence data quality management is vital to its success. All the data included in the project will be from reputable sources.

3 Conclusions

OPTIDRILL aims to develop a drilling advisory system utilising machine learning techniques. In order to achieve that goal, there is a need to establish a data Lifecycle management strategy as part of the process to ensure high quality and reproducible and verifiable data. The OPTIDRILL Data Lifecyle and Curation strategy has been presented in this document, highlighting the methodologies selected for that purpose as well as its implementation throughout the project.