Data Management in Digital Twins for the Oil and Gas Industry: beyond the OSDU Data Platform

Jaqueline B. Correia, Fabrício Rodrigues, Nicolau Santos, Mara Abel, Karin Becker

Federal University of Rio Grande do Sul, Brazil
{jbcorreia, fabricio.rodrigues, nicolau.santos, marabel, karin.becker}@inf.ufrgs.br

Abstract. Competitiveness in the Oil and Gas (O&G) sector has required high technological investments for data-centric decisions. One of the trends is the adoption of Digital Twins (DTs), which use virtual spaces and advanced analytical services to monitor and improve physical spaces. Central to the interconnection of these systems is a Data Fusion Core (DFC) component, which provides data management capabilities. Although the literature has proposed data management functionality in the scope of specific O&G DT applications, different joint efforts towards standardization can be found to deal with data integration and interoperability in the industry. The Open Subsurface Data Universe (OSDU) data platform is an initiative by several partners members of The Open Group consortium created to eliminate data silos in the O&G ecosystem and leverage innovation through a data-driven approach. In this article, we look at the convergence of this effort in providing data management functionalities for digital twins, highlighting strengths, gaps, and opportunities. We investigated the extent to which the OSDU data platform meets the needs of a DFC implementation, with a focus on interoperability, integration, governance, and data lineage. We also propose additional resources for data management in this context, namely data enrichment, workflows, and data lineage. Our main contributions are: (i) analysis of possible data management capabilities for creating a working DFC for an O&G DT and (ii) initial ideas on the complementary role of OSDU data representation and ontologies and how this semantic enrichment can be leveraged in a DFC of a DT.

Categories and Subject Descriptors: H.0 [**Information Systems**]: General; H.2.0 [**Database Management**]: General; H.3.0 [**Information Storage and Retrieval**]: General

Keywords: Data Management, Digital Twin, OSDU, Ontology

1. INTRODUCTION

The Oil & Gas (O&G) industry has leveraged advanced artificial intelligence and Digital Twins (DTs) to create a cycle between the physical and virtual spaces to minimize costs and increase overall productivity. These two spaces can be investigated and explored in a constant feedback process, enabling timely data-centric decisions that add value to the organization.

A DT allows identifying opportunities within an oilfield, such as reducing production costs, increasing operational productivity, monitoring equipment lifetime to increase efficiency, inspecting distinct scenarios through simulations, and others. Existing solutions are very targeted in scope, demonstrating the added value in applications such as early detection of inflows in wells in an offshore environment [Andia and Israel 2018], production optimization [Naufal and Metra 2021], or optimization of drilling operations using historical and real-time data [Dannenhauer et al. 2020]. More encompassing solutions are focused on theoretical issues rather than real implementations [Eilers et al. 2020; Karpov et al. 2021; Brackel et al. 2018], suggesting that DT implementation in the O&G industry is still in early stages [Wanasinghe et al. 2020]. Furthermore, this industry is still very focused on applications that create and manage their data for their particular purposes, with limited scope resulting in silos that make it difficult to exchange data between systems. This siloed ecosystem also affects integrating and enriching data, investigating scenarios using advanced techniques, and controlling the data lineage to

Research supported by CAPES and the PETWIN Project (FINEP financing and LIBRA Consortium).

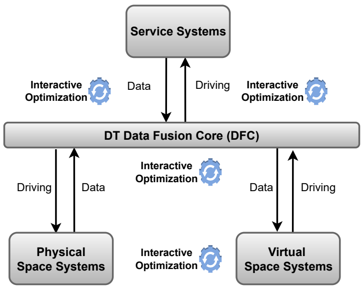

Fig. 1. DT architecture.

trace the path between data sources and decisions.

Different reference DT architectures are proposed in the literature [Wanasinghe et al. 2020]. We adopt the five-component architecture depicted in Figure 1, which highlights the importance of a *Data Fusion Core* (DFC) component. The *Physical Space* component represents the physical assets (e.g., sensors, actuators), while the *Virtual Space* component simulates the physical environment with high fidelity. The *Services* component is related to other enterprise software tools, such as visualization, analytical and predictive resources, model calibration, among others. Each subsystem produces data with intrinsic characteristics in terms of structure, semantics, and systems used to manage their life cycle. They are interconnected through *Connection* components, where the *Data Fusion Core* serves as a point of ingestion of the original data and of return at the right time to direct the process of interactive optimization resulting from the interaction among the subsystems. The DFC role relies on important concepts such as interoperability, data integration, data lineage, and governance.

Different consortia, such as the Professional Petroleum Data Management[1] (PPDM) and Energistics[2], have contributed for data interoperability and integration issues in the O&G industry. They have proposed standards and best practices for data management in the domain (e.g., reference lists, data models) or standardized schemata for exchanging data between systems. These standards provide limited-scope solutions for today's data management challenges, were independently proposed one of another, and are difficult to merge.

In response to these limitations, the Open Group consortium, the O&G industry, cloud service providers, and domain systems providers have been collaborating in the construction of the Open Subsurface Data Universe (OSDU) data platform[3]. This platform focuses on the upstream subsurface scope of O&G. It aims to break down data silos and reduce friction between systems, such that the platform becomes the center of data and the common link for both data-consuming and data-producing systems to drive innovation (e.g., predictive models). This solution will allow building data-centric systems that are authoritative data sources (System of Record - SoR). To this end, it provides an architecture for managing cloud-native data, based on standardized APIs (Application Programming Interfaces) and microservices architecture, scalable and technology-independent. Similar to the concept of Data Lake [Sawadogo and Darmont 2021], it maintains data in its native format, providing resources for contextualization through metadata, as well as resources for governance, physical and logical organization of data, storage and processing of large volumes of data. scalable data. With massive industry support and growing adoption, the OSDU platform has become the *de facto* standard for data management in this domain.

In this article, we critically evaluate the resources offered by the OSDU data platform as a basis for the DFC of a DT focused on the O&G industry, considering the perspectives of data integration, interoperability, governance, and data lineage. We make a case for complementary resources that

[1]http://ppdm.org/
[2]https://www.energistics.org/energistics-university/
[3]https://osduforum.org/

complements the potential of this platform to meet the needs of such a DFC. This article extends our previous position paper [Correia et al. 2021] with an updated related work, a more comprehensive description of the OSDU data platform and discussion on its strengths and limitations, and detailing the complementary roles of the OSDU platform and ontologies for data enrichment in a DFC. Our main contributions are: (i) a critical analysis of data management resources provided by OSDU that can be leveraged to create a functional DFC for an O&G DT and (ii) initial ideas on the complementary role of the OSDU data representation and ontologies in a DFC.

The rest of this text is structured as follows. Section 2 summarizes related work. Section 3 summarizes the aspects of the OSDU data platform that are relevant for our discussions. Section 4 presents a scenario in the O&G domain to illustrate its complexity. Section 5 discusses the potential of the OSDU data platform for the DFC of an O&G-oriented DT, suggesting additional features. Section 6 outlines the complementary role of OSDU data representation and ontology and how this semantic enrichment can be leveraged for DFC functionality. Section 7 presents conclusions and future work.

2. RELATED WORK

Digital twins have a growing importance in different domains, such as smart manufacturing, intelligent health monitoring, building information modeling, and O&G. As an emerging field, it lacks a consensual definition, and different surveys have contributed to a deep understanding of the core concepts, technological challenges, main application areas, and key properties of DTs (e.g. [Barricelli et al. 2019; Jones et al. 2020]). A survey [Wanasinghe et al. 2020] addressing the O&G industry highlights that most existing solutions are targeted at asset integrity monitoring, project planning and life cycle management, drilling, offshore platforms and infrastructure, and intelligent oilfields.

Earlier DTs architectures focused only on connecting a physical system and a mirrored virtual one. Existing DTs for the O&G industry reported in the literature follow this architecture, with *ad hoc* solutions for data management issues such as data extraction and integration of heterogeneous sources, data sanity, data transformation, and data consumption in the virtual environment (e.g., visualization, simulations, event selection) [Sharma et al. 2017; Agarwal and McNeill 2019; Naufal and Metra 2021; Bho 2019; Tang et al. 2018; Andia and Israel 2018]. The existence of data silos and the difficulty of handling multiple, heterogeneous data sources, formats, and types of data are often mentioned as a major difficulty. There are works reporting DTs in the O&G domain propose limited scope solutions for data management problems such as data cleaning [Sha and Zeadally 2015; Andia and Israel 2018], data heterogeneity and integration [Murray et al. 2019; Al-Ismael et al. 2020; Dao et al. 2014; Brackel et al. 2018; Jirkovskỳ et al. 2016] or interoperability [Platenius-Mohr et al. 2019].

Subsequently, a five-component architecture was proposed, which includes the DFC and the Services components [Tao and Zhang 2017]. The DFC acts as a bridge between all the other systems, collecting and consolidating data from separate and heterogeneous systems, to generate driving commands for the other components. In this role, the functions of the DFC can be approximated to the data and knowledge management functionality in Data Lakes [Kronberger et al. 2020]. A Data Lake should provide support for [Sawadogo and Darmont 2021]: (1) storage of large volumes of data in their native format (structured, semi-structured and unstructured data); (2) data search and processing; (3) data life-cycle control; (4) metadata cataloging and management to ensure data quality; (5) implementation of policies and rules for data governance; (6) scalability of storage and processing and (7) an integrated architecture that facilitates data organization for better navigation and data manipulation. It should be pointed out the term "data fusion core" in the DT context refers broadly to all issues related to integrating/merging the data of all other components or spaces of DT [Wang et al. 2019] (physical, virtual, services, and connections), rather than the specific fusion step in the whole data integration process in traditional big data architectures [Dong and Srivastava 2015]. In other words, the DFC addresses the use of all techniques, concepts, tools, and technologies necessary to meet the proper data management for a DT, dealing with significant data management issues such as data integration, interoperability, data lineage, and governance.

Data integration aims to combine data from multiple sources, providing a unified view [Doan et al. 2012]. From the perspective of DTs, data integration plays a critical role in the DFC, which acts as a central point of integration between the physical, virtual entity and the services involved, in an intertwined process [Jones et al. 2020]. The survey in [Gürdür and Asplund 2018] identifies different types of interoperability, where semantic interoperability addresses the meaning of the exchanged digital resources and how their contextual information affects the interaction of entities [Pagano et al. 2013], thus offering semantic aspects to the DFC. Driving the interactions, the DFC also needs to follow the data journey and data lineage (also referred to as data provenance) that focuses on describing where the data came from, the changes made over time, and how it was manipulated [Herschel et al. 2017]. Finally, data governance plays a critical role to the DFC, which, based on their definition, adds with organization, processes, policies, standards, and technologies necessary to manage and ensure that data meets the needs of the business reducing data management costs [Panian 2010].

An ontology is an explicit specification of a conceptualization [Gruber 1993], i.e., it formally specifies the intended meaning of the terms of a vocabulary according to a certain view of the world [Guarino 1998]. When represented in a language with formally defined semantics, such as the Web Ontology Language (OWL) [Grau et al. 2008], semantic reasoners can make automated inferences over its contents (e.g., classifying an individual into a category according to its characteristics) [Mishra and Kumar 2011; Khamparia and Pandey 2017]. Ontologies have been leveraged to solve a number of data management issues in DTs. The use of standard ontologies representing DT entities (e.g., sensors, power plants, manufacturing) is discussed in [Jir 2017] as a means to reduce the semantic heterogeneity, such that the differences/similarities in the modeling of different/similar concepts can be understood regardless differences in modeling. The inter-operation of distinct DTs is addressed in [Platenius-Mohr et al. 2019], where ontologies are deployed to formulate Model Mappings of information that flows between DTs. An ontology-based traceability mechanism is proposed in [Bougdira et al. 2020] with the goal of providing users with a common knowledge base to access and represent data, contributing to their conceptual integration. The correspondence between different standards available for realizing smart manufacturing is addressed in [Grangel-González and Vidal 2021] in the broader context of intelligent manufacturing. In summary, ontologies can provide in the context of DTs an organizing view over the domain that helps professionals of distinct technical profiles to navigate and integrate data from several proveniences.

3. THE OSDU DATA PLATFORM

The OSDU Forum currently has 228 members[4] that contribute to the platform with expertise from different perspectives. The members include players from the O&G industry (e.g. Shell, Chevron, Equinor), cloud service providers (e.g., Amazon AWS, Google, Microsoft), O&G solution providers and standard organizations (e.g., Schlumberger, Halliburton, Cognite, PPDM, Energistics), general software and information technology solutions (e.g., IBM, OSIsoft, RedHat, Dell) and academic institutions (e.g., Federal University of Rio Grande do Sul - UFRGS, University of Oslo).

The OSDU data platform aims to provide a basic core of services for constructing a data-oriented SoR for the O&G industry. Services are cloud-native, provider-agnostic, and must compose applications according to standard protocols and microservices architectures. Cloud service providers implement these services through specific APIs, using their own resources for storage and processing large volumes of data in a scalable, high-performance, and secure way. On the other hand, domain solution providers compete by building solutions based on this platform. The data principles that guided the platform's design define that all data has value and therefore must be globally identifiable, kept in its raw form, and ingested by the system with the least possible friction. In addition, it must be subject to a minimum viable governance process and continually monitored for quality control [The Open Group 2020a]. Forum members (e.g., Schlumberger, Halliburton) have contributed to the construction of

[4]https://www.opengroup.org/osdu/current-members

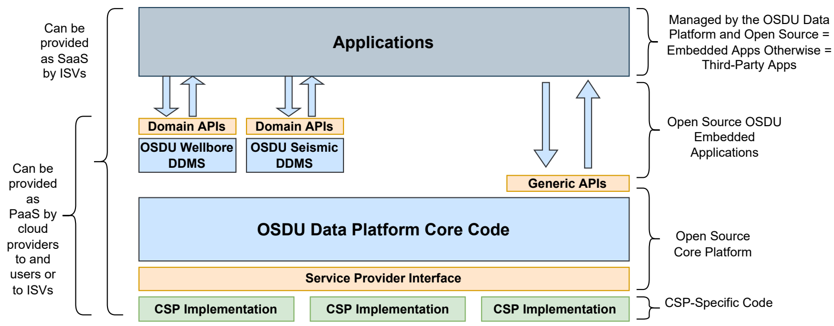

Fig. 2. OSDU Data Platform: big picture. (Adapted from [The Open Group 2020a]).

APIs intended to handle specific data (e.g., seismic data, time series) to promote interoperability of specific solutions and the platform.

Figure 2 shows the big picture of the platform ecosystem. Using microservices' architecture, applications are developed by leveraging generic or specific APIs. The generic APIs provide access to the basic core services of the OSDU data platform, and service providers provide cloud platform-specific implementations for these APIs. The ecosystem also comprises domain-specific services, referred to as *Domain Data Management Services* (DDMS), providing APIs for managing data with specific properties (e.g., Wellbore, Seismic) [The Open Group 2021].

3.1 Architecture and Core Services

Mercury R3 is the denomination of the current version of OSDU platform, and it offers services in the form of APIs aimed at ingesting and storing data, defining metadata schemata, indexing, searching, access control and security, compliance, event notifications, and utility services (spatial references and unit conversions) [The Open Group 2020a]. Figure 3 summarizes the OSDU platform's view about the data life cycle and the services offered by the basic core.

The platform can ingest any type of data, and the ingestion process can be done through both ETL (extract, transform, and load) and ELT (extract, load, and transform) methods. The ETL method moves the data from the source systems into the staging area of the platform, where data undergoes transformations considered necessary. The ELT method enables to ingest from the source system into the platform in its raw format, i.e., without any transformation process. Data must be described by the respective metadata, according to a common data model, detailed in Section 3.2. All metadata is indexed, enabling future search. The indexer/search mechanism is based on Elastic Search. Indexes and indexed documents are saved in a separate persistent store to optimize the search.

To find data of interest, the search service allows one to search full text in string, date range, numeric and geospatial fields. Given that Elastic Search is developed on top of Apache Lucene, queries are formulated according to Lucene syntax, which is based on the use of terms and fields, logical operators, and wildcard operators. The OSDU platform has a pattern for query construction in which the "kind" record field, denoting the corresponding data type, is mandatory since it maps attributes that regulate governance such as authority, origin, partition and namespace. In this way, queries can be built by defining the type of the schema and the properties of interest.

The core also provides services related to governance, which provide the ability to create and manage properties that regulate the legal status of data and how they can be consumed and ingested (compliance). These include authorization of data use, the creation and management of groups and their permissions (entitlements), and definition of the groups of users that regulate data access and usage. The partition service provides the highest level of data isolation within a single OSDU implementation, in which each user is entitled to access data only in the allowed partitions.

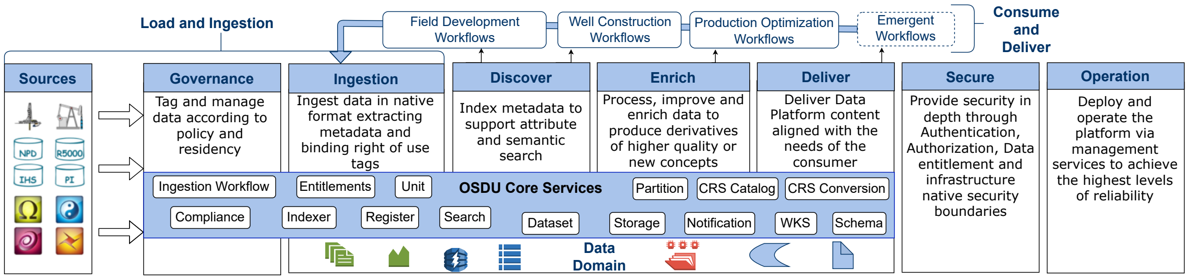

Fig. 3. OSDU Data Platform: data life cycle and basic service core. (Adapted from [The Open Group 2020a]).

Although enrichment is promoted as a universal concept in the platform, the current release provides limited support. It provides a Coordinate Reference Systems catalog (CRS) and supports coordinate transformation required for searching data with a common frame of reference. The Unit service provides a catalog for units of measures and conversion to a common frame of reference.

DDMS components can also utilize the platform core services such as the workflow service for orchestration of processes, the schema service for data definition typically used through on schema write method, and the search service for data search such as a seismic survey [The Open Group 2020b].

3.2 Data Representation

The data model illustrated in Figure 4 shows how information is organized according to OSDU design principles in a data-oriented platform and created, queried and consumed by the different services of the platform. Raw information is stored in the form of data files (*Dataset*), which can be files, sets of files, and in future releases of the platform, databases. All files are contextualized through records (*Record*), in the form of data and metadata. According to the governance principle, each of these records is associated with lists that regulate the access to the information (*Access Control List*), and at least one *Legal Tag*. A Legal Tag is a collection of properties that govern how data can be consumed and ingested, such as unexpired contracts or countries. All records must conform to a schema (*Schema*), which defines the data that all records must contain. The relationship between a record and its corresponding schema is through its type (*kind*). All these elements are described textually in JSON format. There are five types of schemas in OSDU, referred to as *Group Types*: Master Data, Reference Data, Dataset, Work Product Component and Work Product.

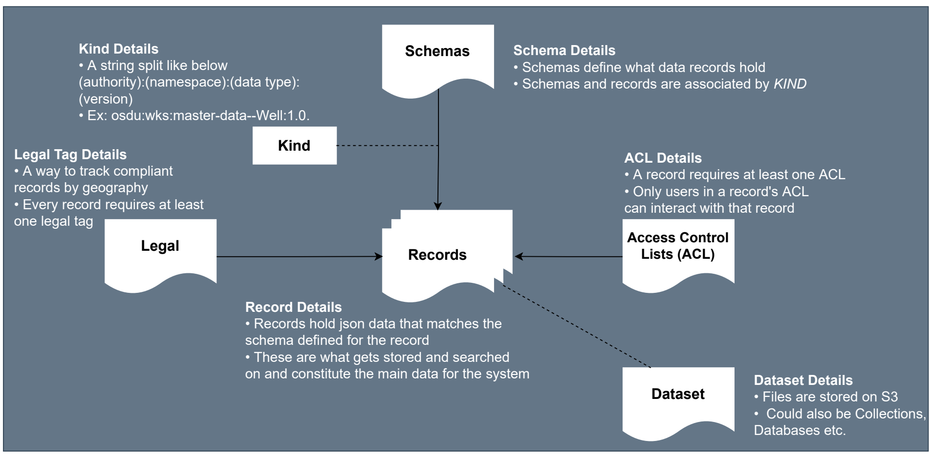

Fig. 4. OSDU simplified Data Model. (Adapted from [Lakshmipathi and Wang 2021]).

A *Master Data* schema is used to describe digital objects that represent physical entities from the real world (e.g., well, wellbore) or business activities (e.g. seismic acquisition). All Master Data objects must have a unique identifier (ID), immutable over time, which can either correspond to a real world identification, or a system assigned ID. All other properties can vary over time, and the changes correspond to the versioning of this object. A Reference Data schema also aims at providing context to objects, but it describes conceptual entities, such as unit measure, domain type, legal status, organization type, tubular component connection type, and others.

A *Dataset* schema provides metadata about digital files, that is, properties that describe information such as file size, format, storage location, and others. It does not describe the file's contents and it is immutable, i.e., whenever the underlying binary file is updated, a new Dataset/File schema should be created. A *Work Product* schema allows expressing, for auditing purposes, the ingestion of a set of data, such as a group of well logs. In this case, the schema refers to a collection of Work Product Component. A *Work Product Component* (WPC) schema allows specifying metadata about the content of a data set or file, allowing enriching it with useful metadata facilitating the search and consumption of data. A WPC is usually the output of a master data and can point to other WPC's as well as to the artifacts (CSV file), and when artifact information is updated, a new WPC should be created. However, it can change the status, purpose, or relationship, and WPC continues with the same identity but with a new version.

The OSDU members work to consolidate OSDU reference schemata for different areas of the O&G domain. The platform currently offers 14 Master Data and 201 Reference Data schemata covering the seismic, wellbore and reservoir areas. Other efforts are currently on place to cover other key domains (e.g., production). The platform also provides a set of canonical and standard schemata for any OSDU implementation, covering key attributes considered relevant to potential platform's customers. The *Schema* service API allows the inclusion of additional attributes in an existing schema using a section called *ExtensionProperties*, or to create new schemata for their own specific needs [The Open Group 2020b]. Governance properties allow defining the properties of a schema as fixed (predefined by OSDU), open (defined by OSDU and companies), or local (organization-specific).

4. ILLUSTRATIVE SCENARIO

In this section, we present a scenario in the O&G context to illustrate its complexity from a data management point. The reservoir simulation on the production phase of an oil field requires important decisions concerning the long-term hydrocarbon extraction from the reservoir. Industry professionals aim to answer questions such as how much oil the field will produce, what methods should be applied for extracting oil and gas from the subsurface, what types of fluid should be injected into the reservoir to maintain good reservoir pressure, how to optimize the field's oil production, and what are satisfying and attainable production targets. It is necessary to simulate petroleum production considering different strategies and scenarios to answer questions such as these. Therefore, production simulation is thus a traditional and essential activity in the O&G industry, and it encompasses a complex set of actions.

The Production Engineer's job is to monitor individual wells or groups of wells to make sure they are producing (or injecting, in the case of injection wells) at their most optimal levels. Downhole equipment and tubing may corrode or erode over time; generated hydrocarbon fluids may deposit waxes or asphaltenes on downhole equipment and tubing; minerals in produced water may cause a scale on downhole equipment and tubing, steel tube may corrode or erode, and so on. All of these factors might have a detrimental impact on the well's performance. Furthermore, equipment and tube diameters ideal at the start of production may be inadequate in depleted reservoir circumstances due to changing reservoir conditions.

An essential step for production optimization is gathering data from various areas, such as flow assurance, reservoir engineering, and equipment maintenance. Therefore, interacting with other divisions in the company is a part of the routine of a production engineer. Production optimization needs a set of real-time time series data such as pressure and temperature values from the downhole and the wellhead. In the case of multiple production zones, pressure and temperature from all the zones must be monitored. Usually, dozens of injecting and producing wells are actively monitored in an offshore oil field to optimize hydrocarbon recovery from the subsurface reservoir.

In summary, production simulation is based on data generated in several other processes and hence depends on excellent data management for its success. Thus, organizations must manage their data, have well-defined governance processes rich in policies and rules that cover the entire data life cycle to improve communication between work teams and facilitate data access and recovery relevant. It must also consider to record all data lineage and develop workflows to automate data management process, minimize the creation of silos by promoting integration and unified visualization of data and enrich the data semantically such that useful data can be identified and explored.

5. DATA FUSION CORE OF A DT FOR O&G INDUSTRY: DISCUSSION

In this section, we evaluate the advantages and limitations of the OSDU platform based on the illustrated scenario presented in Section 4. Our discussion is focused on the needs related to data management in the context of DT, more specifically, aspects of data integration, interoperability, governance, and data lineage.

5.1 OSDU Data Platform: Evaluation

It is clear from Section 3 that the OSDU data platform provides essential resources for a DFC focused on the O&G industry. Considering the scenario illustrated in Section 4, the OSDU platform provides a SoR that is an authoritative data reference for the reservoir simulation, which aids to break the source data silos in this ecosystem. The data generated/consumed by sensors and physical space systems, virtual space simulators, and different services used in the analysis can be all centralized in this platform in their native format and related to the respective contextual information. Core services cover the data management life cycle, from the ingestion of heterogeneous raw data, and record/notification services allow players to monitor events related to their focus of interest. However, it is not yet a complete solution.

In terms of data model, the platform offers a simplified type system and a number of predefined useful schemata. Nevertheless, it does not cover all areas, and in practice, the OSDU forum members need to expend considerable efforts to cover new areas with a consensus modeling. Important directions are the integration of the OSDU model with industry standards. Examples are Energistics O&G interoperability standards (PRODML, RESQML, and WITSML) to handle production, reservoir, and drilling data. Another key standard is CFIHOS[5], which provides support for data from industrial facilities and equipment. These integrations are currently under discussion, but far from concrete results due to their inherent complexity.

The platform compensates for these limitations with many degrees of freedom in defining a new schema or extending existing ones. As each organization is free to define its own schemata, even for the same type of data (or similar), this directly penalizes the semantic interoperability of the data. In a production simulation process involving data from different areas, semantic interoperability is fundamental for a common understanding of the data. In a broader context, creating more specialized schemata on an *ad hoc* basis undermines the exchange of data within a community, as the meaning of a term may have different interpretations by different companies, as well as a distinct set of properties.

Regarding governance, the platform provides an access control mechanism, allowing OSDU customers to manage the people/organizations who can access the data according to their respective level of permissions. It also provides legal metadata (tags), which can be assigned in the schema to make explicit usage rights, contractual obligations, and commercial compliance. The use of these features provides a beneficial organizational structure bringing fluidity to workflows. In the case of

[5]`https://www.jip36-cfihos.org/`

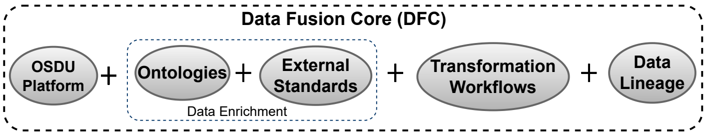

Fig. 5. Data Fusion Core for an O&G DT.

processes that involve the use of data generated at different stages of the upstream chain, such as reservoir simulation, it is essential define which users can/cannot access or edit the data. It also makes explicit the legal and contractual information, bringing transparency in the flow of information.

Although the platform has a powerful search engine, it is limited to syntactical aspects of data schemata. A key value of data integration is identifying data relationships between domains and detecting previously undiscovered patterns. As mentioned in Section 3.1, searches on the platform are performed at the syntactic level, which, despite being very useful, limits the discovery of relevant data since it requires the user to know the properties of the data beforehand. Therefore, improving the data search to the semantic level allows the relevant data to be found even by users who do not know the data deeply, nor the repetition of searches to complete the spectrum of objects possibly represented. It also allows other types of search in terms of semantic properties, semantic relationships, or similarity of concepts. Democratizing the discovery of relevant data is a crucial feature for the OSDU platform ecosystem, as it is for a DFC. In the illustrated scenario, reservoir simulation depends on the search for various fundamental data for the oil and gas production simulations and analysis. A semantic metadata model based on ontology would allow the semantic level, an essential resource for the engineer to search for similar or related documents, for instance, to retrieve and identify production wells (i.e., wells used to drain subsurface oil and gas) with similar information.

The platform is limited to primary data lineage or provenance processes. Its schemata provide metadata that allows tracking the data sources, the user who created and updated the object instance, creation and alteration date and time, legal properties that define the ancestry of relationships based on the legal contract, and finally, allow the schema versioning keeping all versions available for possible later checks. Although this metadata is very relevant, it does not allow to keep track of the decisions this data generated, in which analysis and simulation models this data was used, the transformations performed in the data to input the models, which applications consumed the data and why, among others. In a simulation process, this information allows one to deeply understand the generated results and improve them to achieve even more reliable and accurate production forecasts.

In summary, we conclude that the OSDU platform provides a comprehensive set of resources for data management and which can be the basis of a DFC in a DT. It integrates all data and puts it at the center by reducing data silos, has an architecture that enables services' extensibility, and provides metadata for governance and limited data lineage. Despite the identified limitations, we deem it as a good starting point for the services it provides and the standard that it represents in this industry. In this way, one can take advantage of all available data management solutions it provides and leverage these foundations to provide resources for improving data integration, governance, lineage data, and interoperability to obtain a useful DFC for an O&G industry Digital Twin.

5.2 Complementary Resources for the DFC

Figure 5 suggests some complementary resources aimed at managing data in a DFC. The OSDU platform allows the creation of applications that can run within the OSDU platform as an additional service. These new applications can also be used by internal workflow orchestration services, or these new applications can connect to the platform via APIs and run externally to it, according to the OSDU ecosystem (Figure 2). Therefore, the complementary resources shown in Figure 5 could be added to the OSDU platform in both manners.

Data enrichment. Enriching data means using resources to make them rich in context and semantics, aiming at a common understanding. In this sense, existing ontologies that describe different aspects of the O&G domain can be used to add semantic value to the syntactic metadata embedded in the OSDU data platform. Existing consolidated ontologies can be leveraged for this purpose, such as SOSA (Sensor, Observation, Sample, and Actuator) and SSN (Semantic Sensor Network)[6] from the IoT and manufacturing domains, or OntoCAPE and ISO 15926, from the petroleum industry. External standards used by the O&G industry (e.g. PPDM, Energistics, CFIHOS) can also be applied to add value to meta descriptions. Several functionalities can be based on semantic enrichment, among them [Xiao et al. 2018; Schneider and Šimkus 2020]: a) expansion of the search service to semantic characteristics (e.g., by similarity between concepts, by related properties); b) enrich the data transformation and lineage process with semantic metadata; c) domain inferences, based on prior knowledge; etc. In the scenario illustrated in Section 4, the engineer could reuse data from other assets (e.g., in similar wells); allow correlations between similar wells based on some characteristic; evaluate experiments/simulations done with similar purposes; among others.

Transformation Workflows. Data transformation tasks are currently performed by generic or specific purpose data transformation programs or scripts. Iterative data transformation services allow the creation of a transformations workflow, where a user can, from a predefined set of operations, apply transformations on a collection of data, as well as create different sets of data transformation operations [Omitola et al. 2012]. The use of services or iterative data transformation ontologies brings scalability to data transformation. In the scenario illustrated in Section 4, this resource would facilitate the reservoir engineer's job to, for example, select subsets of interest, automatically perform the conversion of units, adjust temporal granularity, or integrate data from different data sets. They must be accompanied by resources that associate metadata to the transformation history from their original sources, through provenance mechanisms [Simmhan et al. 2005].

Data lineage. Often used synonymously with provenance, data lineage capabilities also include the business context in terms of origins and decision. It is essential for the transparency not only of the data origin but also of the entire flow of transformations to which they were submitted and the results/decisions they generated. Through the lineage, it must be possible to follow the quality and reliability of the data, audit data traces, allowing the replication of procedures, assigning properties or responsibilities (e.g., copyright, error), or providing informational context that can be consulted and analyzed [Simmhan et al. 2005]. In a system like DT, in which the decisions that feedback and adjust the different spaces, physical and virtual, must also be the object of the lineage [Eirinakis et al. 2020]. In the illustrative scenario, the engineer needs to keep track of all the simulations/experiments carried out, their relation to the explored data, and their effects on the explored spaces.

In the following section, we focus on the task of fetching production data that is part of the scope of the production optimization and simulation process described in Section 4. We represent a production time series using the OSDU platform schemata. We then similarly use a small ontology to show the semantic benefits that could be obtained through this resource.

6. COMPLEMENTING OSDU WITH AN ONTOLOGY: AN EXAMPLE

A relevant portion of historical production data used in the O&G industry comes in the form of production time series. Production simulation depends on these time series to assist in decision-making regarding an oil field's economic and operational viability. Hence, finding proper time series and dealing with them significantly impacts petroleum production optimization and related processes that use this data. Let us consider the following hypothetical situation:

– There are plenty of data in the form of time series gathering the result of periodical measurements of certain properties related to wellbores, such as bottomhole pressure and temperature (i.e., pressure and temperature at the bottom of a wellbore);

[6]https://www.w3.org/TR/vocab-ssn/

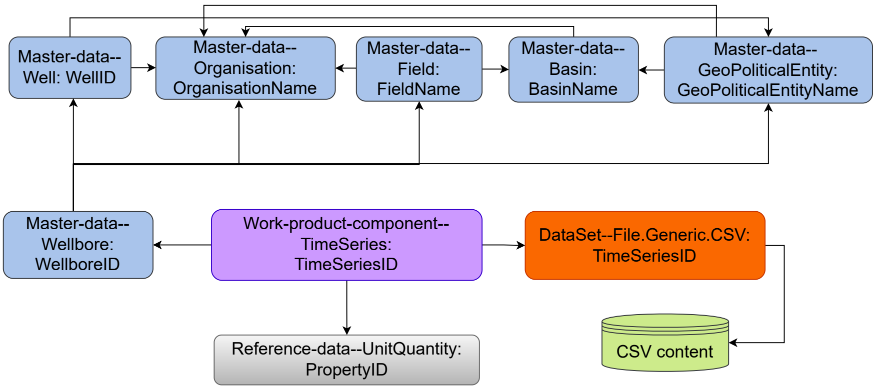

Fig. 6. Representation of the OSDU data schemata for an O&G time series.

– Each time series is stored in a CSV (comma-separated values) file, in which each row consists in the timestamp when the property was measured and the corresponding measured value. These files were transferred to the OSDU platform and related to the appropriate metadata records;
– To find the production time series that are stored in the OSDU platform and use them in the production monitoring and optimization process, the production engineer must be able to find the time series corresponding to properties such as bottomhole temperature and bottomhole pressure, and several other types of related time series.

Figure 6 outlines the main OSDU schemata used to represent and contextualize the bottomhole pressure and bottomhole temperature time series in the simple context described above. The time series data files in their native format are depicted as a green cylinder. The OSDU schemata define the structure of the metadata records that provide computational and business meaning to these datafiles. The blue boxes represent the OSDU *Master Data* schemata about the domain of specific objects, hence relating a time series to a wellbore entity, which in turn is related to other O&G entities. Arrows depict references, which are made using typed attributes. The purple box represents a *Work Product Component* (WPC) schema to represent metadata about business-oriented meaning of a time series of interest. It contains an attribute that defines the unit the time series refers to, represented by a *Reference Data* schema (gray box). In this way, it is possible to describe WPC time series records related to a specific wellbore and some property (e.g., downhole pressure, downhole temperature). The orange box represents the OSDU *Dataset* schema to represent metadata about computational-related aspects of the time series files (e.g., CVS file format, size, storage location).

Considering these schemata, Figure 7 presents a JSON excerpt from a Time Series WPC record (*kind*) which describes a time series dataset. The property `UnitQuantityID` contains a Unit Quantity record id specifying that the time series refer to downhole pressure. The `WellboreID` attribute stores a record id for a specific wellbore from which these time series datasets were measured. Finally, the `Datasets` attribute stores a list containing references to the *Dataset* records describing the computational-related aspects of the time series data CSV files.

The OSDU platform enables making valuables syntactic queries. Figure 8 shows two queries nested in JSON syntax, considering the OSDU representation for time series and our hypothetical scenario. The search structure follows a pattern[7], where the keyword `kind` defines the type of schema, partition, or data authority, and `query` defines which properties will be retrieved and under which conditions. The keyword `nested` allows nested queries that navigate at different levels of depth within a JSON object. The query in Figure 8-(a) retrieves all records of `kind` time series WPCs that have the downhole pressure property with a certain value. Figure 8-(b) shows a slightly different query, which

[7]A detailed description of the search possibilities and syntax can be found at https://community.opengroup.org/osdu/platform/system /search-service/-/blob/master/docs/tutorial/SearchService.md

```
{
        "id": "namespace:work-product--component--TimeSeries:lqa9ol2ws",
        "kind": "osdu:wks:work-product-component--TimeSeries:1.0.0",
        ...

        "data": {

                ...

                "WellboreID": "namespace:master-data--Wellbore:wellbore-xyz",
                "properties": [
                        {
                                "UnitQuantityID": "namespace:reference-data--UnitQuantity:bottomhole_pressure:0ok3wa5rd"
                        }
                ],
                "Datasets": [
                        "namespace:dataset--File.Generic.CSV:abc123wz-p0o9-i8u7-y6t5-r4e3w2q1mkon",
                ]

                ...
        }
}
```

Fig. 7. Snippet of a WPC time series record for a bottomhole pressure time series.

(a)

```
{
  "kind" : "osdu:wks:work-product-component--TimeSeries:1.0.0",
  "query":"nested(
        data.properties,
        UnitQuantityID:\"namespace:reference-data--UnitQuantity:bottomhole_pressure:0o3wa5rd\")"
}
```

(b)

```
{
  "kind" : "osdu:wks:work-product-component--TimeSeries:1.0.0",
  "query":"data.WellboreID:\"namespace:master-data--Wellbore:wellbore-xyz\" AND
   nested(
        data.UnitQuantityID,
        \"namespace:reference-data--UnitQuantity:bottomhole_pressure:0ok3wa5rd\" OR
        \"namespace:reference-data--UnitQuantity:bottomhole_temperature:iam9s3nb\"
)"
}
```

Fig. 8. Examples of queries using the OSDU platform.

aims to retrieve time series `kind` WPC records that refer to a specific *Master Data* and that either have specific values either in the background pressure property or the downhole temperature property.

However, an engineer would not be able to retrieve other time series, such as pressure in other parts of the wellbore (wellhead, bottomhole, choke), other fluid types (gas, oil, water), or other physical properties (temperature) in a single query. In other words, the presented metadata structure cannot cover the semantics of the terms underlying the queries.

Information Artifact Ontology (IAO)[8] is a well-known ontology widely deployed to described time series. Hence, time series related to oil production monitoring could be instances of the category `Time Sampled Measurement Data Set` from IAO. This category makes explicit the conceptualization of time series using other categories, such timed measurement data, the scalar measurement data that classifies the variable (e.g., pressure, temperature) and a label identifying the unit of measurement (e.g., psi, bar, °C, °F). Figure 9 describes an excerpt of the IAO for time sampled datasets specialized for petroleum data. It includes the main categories employed in our hypothetical situation (i.e., wellbore, bottomhole pressure, bottomhole temperature), using a well-established top-level ontology

[8]https://obofoundry.org/ontology/iao.html

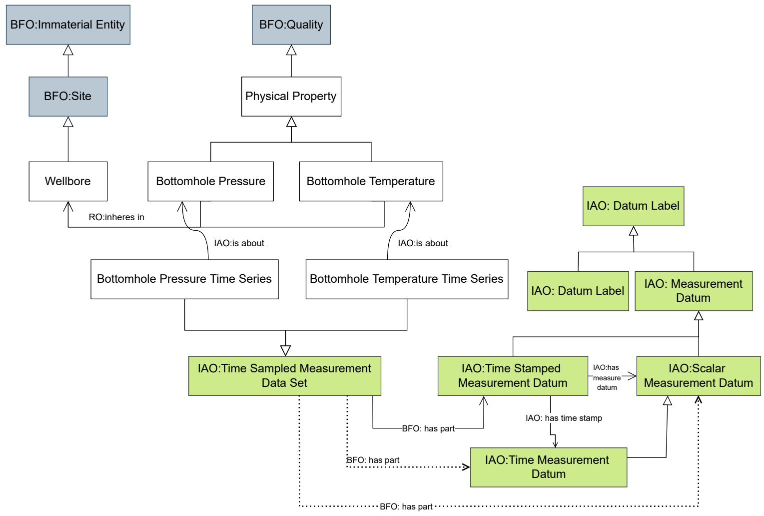

Fig. 9. Excerpt from the IAO Ontology.

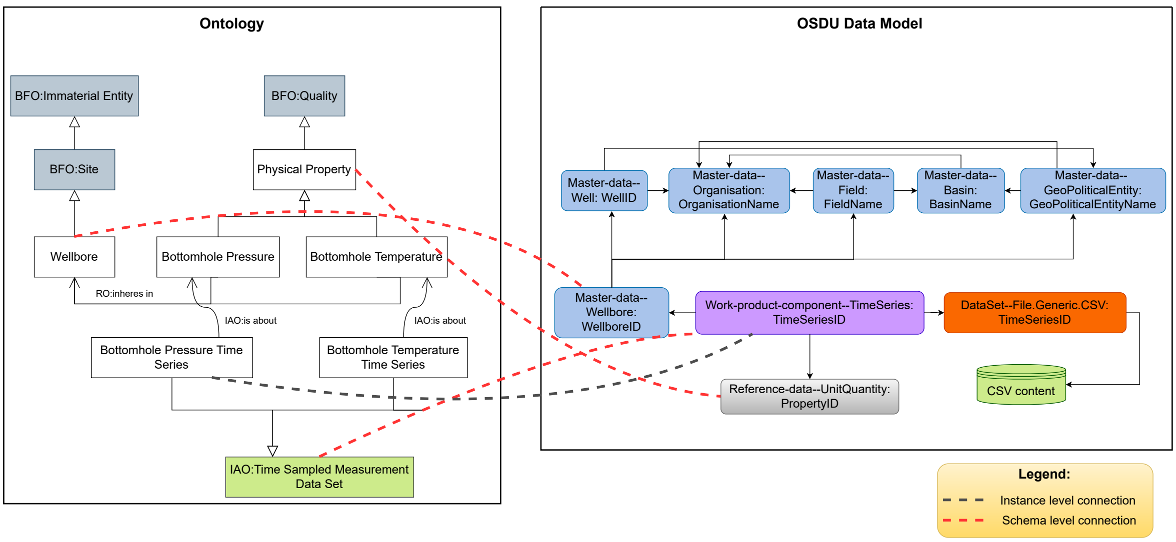

Fig. 10. Complementarity between ontology and OSDU data model.

BFO [Arp et al. 2015] to specialize the concepts, the same top-level ontology used in IAO. This example ontology defines a general class for the already mentioned properties, bottomhole pressure and bottomhole temperature, namely *Physical property*, i.e., a quality that corresponds to some aspect of the physical state of an object. Then, both *bottomhole pressure* and *bottomhole temperature* are defined as sub-types of *physical property*. Hence, we can provide additional semantics for the OSDU records employing this ontology.

Assuming this ontology is described in some standard notation (e.g., OWL), we could make links between the OSDU records to semantically enrich them, and leverage this enrichment for several purposes. A simple approach would be to annotate each OSDU record along with an `ontologyClass` attribute, to store the name of the ontology class that corresponds to the type of time series described in the file. For instance, considering the Time Series WPC schema, we could include a property `ontologyClass` to related it to the concept `IAO:Time Sampled Measurement Dataset`. Hence each record (e.g. Figure 7) could be related not only to the specific OSDU schema (*kind*), but also to ontological concepts (`ontologyClass:"BottomholePressureTimeSeries"`). Then, we could create a knowledge base using the proposed ontology and populate it with an individual for each component of the work product that describes a time series. Each individual would be named according to the contents of the `id` attribute of the file and would be an instance of the class referred to by `ontologyClass`. Figure 10 shows the existing correspondences between the ontology and the OSDU data model in relation to the concept of well and time series at the schema and record levels. While the former approximates the schema to a particular conceptualization or class concept (e.g., well entity, physical property, and time series), and the latter refines it with specific categories for a given record, i.e., it is the class instance (for example, downhole pressure time series of a given wellbore "XYZ").

Based on this enrichment, there are several possible applications. For instance, we could develop semantic queries to search for time series of physical properties in general. So using the DL Query Tab[9] in Protégé, we could run a simple DL Query like

```
'Time sampling measurement dataset' and 'is about' some 'physical property'
```

and retrieve all (and only) instances corresponding to the time series on physical properties. With these aspects, it is possible to search for time series with different structures, allowing to approximate a time series of pressure with scalar measurement data of other variables using physical properties. Alternatively, we could do the same programmatically by using tools such as the OWL API[10].

We exemplified the complementary of a simple ontology for O&G production time series search within the OSDU platform. Following this same line of thinking, the addition of semantics to the huge amount of data that is produced in an offshore enterprise is a fundamental step towards the employment of a data fusion core of a DT. The ontological enrichment makes explicit concepts that underlie OSDU schemas, and which could be leveraged for other purposes. For instance, it could be leveraged to validate records of a given OSDU schema in term of the coherence of the referred objects (e.g. a bottomhole pressure time series must measure pressure and the unit in which it is measured must be compatible). In transformation workflows, it can also be used to maintain the semantic meaning of the entities (e.g. a conversion of a pressure measured in some unit of measurement). It can also be leveraged for data traceability, from source, to modeling and respective decisions.

7. CONCLUSION AND FUTURE WORK

This article describes the importance of a DFC for a DT aimed at the O&G industry. The satisfactory functioning of a DFC requires dealing with aspects of data integration, interoperability, governance, and data lineage. Based on these needs and a common scenario in the O&G industry, we evaluated the advantages and limitations of the OSDU data platform. We concluded that the OSDU platform is an excellent starting point for creating a successful DFC, but complementary resources are required to semantically enrich the data with ontologies and standards for more advanced functionality, transformation flows, and data lineage. To better exemplify the complementarity of the OSDU platform with other resources, we illustrated a hypothetical situation that involves the search for time series within the OSDU platform and presents the benefits that semantic enrichment with a simple ontology can bring. For future work, we intend to extend the ontology shown here and develop a proof of concept that integrates the ontology to the services of the OSDU platform. We have advanced a basic mechanism to relate OSDU schemas/records to ontologies, which should be further investigated and assessed. Other ontologies can be considered in the future, such as the Semantic Sensor Network (SSN), which has become a standard in industry 4.0. Adding an ontology such as the SSN would bring important information for enriching the OSDU schemas by associating the data with the sensor, allowing identification of the sensor that generated the data and the acquisition method, estimating the measurement reliability through the equipment's history, or identifying discrepancies between values of the same property measured by different sensors. In addition to semantic search, other applications

[9]protegewiki.stanford.edu/wiki/DLQueryTab
[10]github.com/owlcs/owlapi/

are to leverage semantic enrichment to automatically generate OSDU records, validate records, and use the semantic definitions for traceability and transformation mechanisms.

REFERENCES

Understanding data heterogeneity in the context of cyber-physical systems integration. *IEEE Transactions on Industrial Informatics* 13 (2): 660–667, apr, 2017.

Digital Twin of Subsea Pipelines: Conceptual Design Integrating IoT, Machine Learning and Data Analytics. OTC Offshore Technology Conference, vol. Day 1 Mon, May 06, 2019, 2019. D011S010R004.

Agarwal, P. and McNeill, S. Real-time cleaning of time-series data for a floating system digital twin. In *Offshore Technology Conference.* OnePetro, 2019.

Al-Ismael, M., Al-Turki, A., and Al-Darrab, A. Reservoir simulation well data exchange towards digital transformation and live earth models. In *International Petroleum Technology Conference.* OnePetro, 2020.

Andia, P. and Israel, R. A cyber-physical approach to early kick detection. In *Proc. of the (IADC/SPE) Drilling Conference and Exhibition*, 2018.

Arp, R., Smith, B., and Spear, A. D. *Building Ontologies with Basic Formal Ontology.* MIT Press, 2015.

Barricelli, B. R., Casiraghi, E., and Fogli, D. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access* vol. 7, pp. 167653–167671, 2019.

Bougdira, A., Akharraz, I., and Ahaitouf, A. A traceability proposal for industry 4.0. *Journal of Ambient Intelligence and Humanized Computing* 11 (8): 3355–3369, aug, 2020.

Brackel, H.-U., Macpherson, J., Mieting, R., and Wassermann, I. An open approach to drilling systems automation. In *SPE Asia Pacific Oil and Gas Conference and Exhibition.* OnePetro, 2018.

Correia, J., Abel, M., and Becker, K. Núcleo de fusão de dados de um gêmeo digital da indústria de petróleo e gás. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados.* SBC, Porto Alegre, RS, Brasil, pp. 343–348, 2021.

Dannenhauer, C. E., Bastos Baptista, G. L., Szwarcman, D., Martins da Silva, R., and Martins Plucenio, D. Real-time physical models with learning feedback as a digital twin architecture. In *Offshore Technology Conference.* OnePetro, 2020.

Dao, M.-S., Pongpaichet, S., Jalali, L., Kim, K., Jain, R., and Zettsu, K. A real-time complex event discovery platform for cyber-physical-social systems. In *Proceedings of International Conference on Multimedia Retrieval.* pp. 201–208, 2014.

Doan, A., Halevy, A., and Ives, Z. *Principles of data integration.* Elsevier, 2012.

Dong, X. L. and Srivastava, D. *Big Data Integration.* Morgan Claypool Publishers, 2015.

Eilers, J., Jacquin, S., and Burgos Basile, N. Asset reliability in practice: An effective combination of apm, ai and simulation solutions within a one stop environment. In *Abu Dhabi International Petroleum Exhibition & Conference.* OnePetro, 2020.

Eirinakis et al., P. Enhancing cognition for digital twins. In *2020 IEEE Intl. Conf. on Engineering, Technology and Innovation (ICE/ITMC)*, 2020.

Grangel-González, I. and Vidal, M. E. Analyzing a Knowledge Graph of Industry 4.0 Standards. In *The Web Conference 2021 - Companion of the World Wide Web Conference, WWW 2021.* Association for Computing Machinery, Inc, pp. 16–25, 2021.

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., and Sattler, U. Owl 2: The next step for owl. *Journal of Web Semantics* 6 (4): 309–322, 2008.

Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5 (2): 199–220, 1993.

Guarino, N. Formal ontology and information systems. In *Proceedings of the International Conference on Formal Ontology and Information Systems (FOIS'98).* IOS Press, Amsterdam, Netherlands, pp. 3–15, 1998.

Gürdür, D. and Asplund, F. A systematic review to merge discourses: Interoperability, integration and cyber-physical systems. *Journal of Industrial Information Integration* vol. 9, pp. 14 – 23, 2018.

Herschel, M., Diestelkämper, R., and Lahmar, H. B. A survey on provenance: What for? what form? what from? *The VLDB Journal* 26 (6): 881–906, 2017.

Jirkovskỳ, V., Obitko, M., and Mařík, V. Understanding data heterogeneity in the context of cyber-physical systems integration. *IEEE Transactions on Industrial Informatics* 13 (2): 660–667, 2016.

Jones, D., Snider, C., Nassehi, A., Yon, J., and Hicks, B. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology* vol. 29, pp. 36–52, 2020.

Karpov, R. B., Zubkov, D. Y., Murlaev, A. V., and Valiullin, K. B. Drilling performance and data quality control with live digital twin. In *SPE Russian Petroleum Technology Conference.* OnePetro, 2021.

Khamparia, A. and Pandey, B. Comprehensive analysis of semantic web reasoners and tools: a survey. *Education and Information Technologies* 22 (6): 3121–3145, 2017.

Kronberger, P., Dabrowski, P., Chacon, J., and Bangert, P. The Digitalization Journey of the Brage Digital Twin. SPE Norway Subsurface Conference, vol. Day 2 Tue, November 03, 2020, 2020.

Lakshmipathi, K. and Wang, K. Energy data platform - application developer bootcamp. Tech. rep., 2021.

Mishra, R. B. and Kumar, S. Semantic web reasoners and languages. *Artificial Intelligence Review* 35 (4): 339–368, 2011.

Murray, P., Wattis, Z., Bain, B., Golowczynski, M., and Sadd, A. Towards a digital twin supporting risk based decision making for offshore installations. In *SPE Offshore Europe Conference and Exhibition*. OnePetro, 2019.

Naufal, A. A. and Metra, S. A digital oilfield comprehensive study: Automated intelligent production network optimization. In *SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition*. OnePetro, 2021.

Omitola et al., T. Capturing interactive data transformation operations using provenance workflows. In *Extended Semantic Web Conference*. Springer, pp. 29–42, 2012.

Pagano, P., Candela, L., and Castelli, D. Data interoperability. *Data Science Journal* vol. 12, pp. GRDI19–GRDI25, 2013.

Panian, Z. Some practical experiences in data governance. *World Academy of Science, Engineering and Technology* 62 (1): 939–946, 2010.

Platenius-Mohr, M., Malakuti, S., Grüner, S., and Goldschmidt, T. Interoperable digital twins in iiot systems by transformation of information models: A case study with asset administration shell. In *Proceedings of the 9th International Conference on the Internet of Things*. pp. 1–8, 2019.

Platenius-Mohr, M., Malakuti, S., Grüner, S., and Goldschmidt, T. Interoperable digital twins in IIoT systems by transformation of information models: A case study with asset administration shell. In *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. ICST, 2019.

Sawadogo, P. and Darmont, J. On data lake architectures and metadata management. *J. Intell. Inf. Syst.* 56 (1): 97–120, 2021.

Schneider, T. and Šimkus, M. Ontologies and Data Management: A Brief Survey. *KI - Künstliche Intelligenz* 34 (3): 329–353, 2020.

Sha, K. and Zeadally, S. Data quality challenges in cyber-physical systems. *Journal of Data and Information Quality (JDIQ)* 6 (2-3): 1–4, 2015.

Sharma, P., Hamedifar, H., Brown, A., and Green, R. The dawn of the new age of the industrial internet and how it can radically transform the offshore oil and gas industry. In *Offshore Technology Conference*. OnePetro, 2017.

Simmhan, Y. L., Plale, B., and Gannon, D. A survey of data provenance in e-science. *SIGMOD Rec.* 34 (3): 31–36, Sept., 2005.

Tang, S., Wang, R., Zhao, X., and Nie, X. Building Cloud Services for Monitoring Offshore Equipment and Operators. OTC Offshore Technology Conference, vol. Day 1 Mon, April 30, 2018, 2018. D011S011R005.

Tao, F. and Zhang, M. Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing. *Ieee Access* vol. 5, pp. 20418–20427, 2017.

The Open Group. The open group guide osdu™ system concept. Tech. rep., 2020a.

The Open Group. The Open Group Guide OSDU™ Reference Architecture. Tech. rep., 2020b.

The Open Group. OSDU™ Operator Data Loading Guide. Tech. rep., 2021.

Wanasinghe et al., T. Digital twin for the oil and gas industry: Overview, research trends, opportunities, and challenges. *IEEE Access* vol. 8, pp. 104175–104197, 2020.

Wang, P., Yang, L. T., Li, J., Chen, J., and Hu, S. Data fusion in cyber-physical-social systems: State-of-the-art and perspectives. *Information Fusion* vol. 51, pp. 42–57, 2019.

Xiao et al., G. Ontology-based data access: A survey. In *Proc. of the Twenty-Seventh Intl. Conf. on Artificial Intelligence, IJCAI-18*. pp. 5511–5519, 2018.