UNIVERSITY
OF OSLO

Master Thesis

Diving Into The Subsurface: Disambiguating Industrial Data

A Case Study on Data Contextualisation in The Subsurface Domain

Stefan Spanic

Eirik Borgen Egge

Informatics: Digital Economics and Leadership
30 ECTS

Department of Informatics
Faculty of Mathematics and Natural Sciences

20th May 2026

Stefan Spanic

Eirik Borgen Egge

Diving Into The Subsurface: Disambiguating Industrial Data

A Case Study on Data Contextualisation in The Subsurface Domain

Supervisor:
Dragana Paparova

Abstract

This thesis examines how industrial data are disambiguated through contextualisation as they depart from their origin and are reused in other settings. We address this through a qualitative single-case study of Equinor, a Norwegian oil and gas operator transitioning parts of its subsurface workflow onto the Open Subsurface Data Universe (OSDU), drawing on nine semi-structured interviews triangulated with documentary material. Findings show the work that constitutes contextualisation, and how the transition to OSDU reorganises this work rather than replacing it. We develop a process model with three subprocesses — *categorising, detaching, and reconstructing* — through which disambiguation proceeds non-linearly: ambiguity is stabilised as embedded data are categorized into shared models, reintroduced as data detach from their source systems, and subsequently addressed as practitioners re-fit detached data to new purposes. The model extends the IS data work literature and locates contextualisation for AI upstream in the preparation of data, rather than only at the point where outputs are consumed.

Keywords: Data Contextualisation, Disambiguation, Data Work, AI, Subsurface Data, Industrial Data, OSDU, Oil and Gas

Acknowledgements

We would like to begin by thanking our supervisor, Dragana Paparova, for her excellent feedback and guidance in writing this master's thesis. Your insights and encouragement have been greatly appreciated, and your professionalism has shaped how we approach our own work. We are grateful for everything we have learned from you along the way.

We would further like to thank our informants who have dedicated their time to sharing their knowledge with enthusiasm. A special thanks goes to Einar Landre for facilitating the project and helping us connect with relevant interviewees.

This master's thesis would not have been the same without support from all of our wonderful classmates at Digøk. Thank you all for the great conversations, lunches on the sunny balcony, and much needed walking breaks.

Finally, we want to thank each other for challenging and encouraging one another, and for the company through all of it. Whether mapping out big ideas on whiteboards in cramped meeting rooms or working through problems over lunch, we have learned a great deal from one another.

It was demanding and rarely straightforward, but we enjoyed it, and we are leaving better equipped than we started.

Regards,

Eirik Borgen Egge & Stefan Spanic

Declaration of Generative Artificial Intelligence usage

In this scientific work, generative artificial intelligence (AI) has been used. All data and personal information have been processed in accordance with the University of Oslo's regulations, and we, as the authors of the document, take full responsibility for its content, claims, and references. An overview of the use of generative AI is provided below.

Generative AI tools have been helpful in this process. That being said, we have taken great care to minimise their use and only use them supplementally for particular tasks.

In our text editor of choice, Overleaf, there is an AI-based, grammar– and spelling corrections tool called Writefull. It has at times been used to provide minor spelling and punctuation suggestions.

GPT UiO is a service developed by UiO that makes it possible to use OpenAI's GPT models within UiO's privacy requirements. This tool has been used to suggest translations and refinements in cases where the prompts contain data from the thesis. We have minimised the use of these instances and tried to translate to the best of our abilities by ourselves, where possible.

Due to their performance and our familiarity with the tools, Claude's Sonnet 4.6, developed by Anthropic and ChatGPT's GPT-5, developed by OpenAI, have also been used. In the cases where these tools were used, the prompts did not include data from the thesis. The tools were used for language support, such as refining or rephrasing certain sentences, as well as inspiration for how to structure certain parts of the thesis. For example, a prompt could look like "How is a research background typically structured in IS?" or "Offer alternatives to this phrasing: 'traditional ways of conceiving data'". Perhaps most usefully, the tools have been used to support LaTeX debugging and structuring the various tables and figures in the thesis. Where the tool has been used in this case, we have used filler-text in place of actual data to help with formatting.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The push to make industrial operations more 'data-driven' (Fischer et al., 2023) has intensified expectations of the role of data in industrial work. Organisations in heavy-asset industries are under increasing pressure to extract more analytical value from the data they already hold, and to do so faster and more consistently than before. Two narratives have shaped how the industry addresses the need to become more data-driven. The first is that data — commonly referred to as 'the new oil' (Economist, 2017) — is an underused valuable asset. The second is that artificial intelligence (AI) will provide the analytical leverage needed to generate value from data assets (Davenport, 2018). Both presume that once relevant data are consolidated and made accessible, they can be reused outside the context in which they were produced, and reshape how industrial organisations operate.

This presumption is what a substantial body of information systems (IS) research has problematised. Studies of data work have shown that data are not inherently valuable in the way the narrative implies. They are produced through situated practices and remain interpretable only when accompanied by the tools, practices, and tacit knowledge that produced them (Mikalsen & Monteiro, 2021; Østerlie & Monteiro, 2020; Parmiggiani et al., 2022). Another literature stream within IS research has shown similar insights related to AI: model outputs only become actionable through the work of human experts who can read those outputs against

the situations they are meant to inform (Lebovitz et al., 2021; van den Broek et al., 2021; Waardenburg & Sergeeva, 2021). Both strands indirectly touch on a process this thesis calls *data contextualisation*: the work of producing data alongside the conditions required to interpret them, and restoring those conditions when the data enter settings other than the ones that produced them.

Working within this body of research, this literature has shown that data carry the residue of where and how they were produced, and that they become more ambiguous as they are distanced from that origin and reused elsewhere (Aaltonen et al., 2021; Mikalsen & Monteiro, 2021; Østerlie & Monteiro, 2020). This thesis examines from a different angle how that ambiguity is worked through in practice: what is done to keep data interpretable as they move, and where in that movement the work falls, particularly once the original practitioners and applications are no longer at hand. Within the AI strand, attention has fallen mostly on the point where outputs are interpreted and made trustworthy for use, rather than on the data side, where the conditions for interpretation are set before any analysis takes place. The thesis takes this up through the following research question (RQ):

RQ: *How is industrial data disambiguated through contextualisation as it departs from its origin and is reused in other settings?*

The thesis addresses this through an in-depth study of Equinor, an oil and gas operator on the Norwegian Continental Shelf. Equinor's core work is recovering hydrocarbons from rock several kilometres beneath the seabed. Direct access to that rock is sparse, limited to samples brought up from the few wells drilled into any given formation. The rest is inferred from seismic surveys at the surface and from measurements taken inside the wells as they are drilled. The subsurface exists for the operator largely as data, and what is known about it is shaped by how it is known (Monteiro, 2022, pp. 15, 21). Interpretation runs through every part of that picture, and so does ambiguity.

Within Equinor, the production and use of these data has historically been organised around proprietary applications used by domain specialists. Inside those

applications, the data and the interpretive work bound up with them have lived together. That arrangement is changing, slowly and unevenly. For some parts of the work, Equinor has been adopting the Open Subsurface Data Universe (OSDU), a standardised data platform developed under The Open Group. Adoption has been partial. Established applications and workflows remain in use, and OSDU sits alongside them rather than replacing them. The established way of working and the gradual reorganisation around the platform are visible at once, which lets us examine how contextualisation changes as subsurface data move out of the environments that have historically held them.

The study is a qualitative case study (Gerring, 2004), following an interpretive research design (Walsham, 1995). The empirical material consists of nine semi-structured interviews conducted between January and April 2026 with informants across petrophysics, geology, data engineering, data science, and IT, triangulated with company documents, regulatory material, and OSDU documentation. The analysis uses elements from the Gioia methodology (Gioia et al., 2013) combined with some categorising and connecting strategies (Maxwell & Miller, 2008). The latter were used to trace how the empirical material related across the subsurface workflow, thereby preserving the processual sequence and revealing how context and ambiguity evolve as data move across contexts.

Our findings show how subsurface data are disambiguated through data contextualisation: the process through which industrial data become analytically usable by producing the data together with the conditions needed to interpret them, and reassembling those conditions when the data are reused in a setting other than the one that produced them. In our case, the transition to OSDU reorganised this work rather than removing the need for it. From these insights, we develop a process model of how industrial data are disambiguated through contextualisation, and how that work is reorganised when data move onto a shared platform. The model is grounded in the subsurface case but holds for domains where data carry the residue of substantial prior interpretive work and lack a directly accessible referent.

This thesis contributes to IS literature on data work by defining data contextualisation as the process through which data become analytically valuable (Mikalsen & Monteiro, 2021; Østerlie & Monteiro, 2020; Parmiggiani et al., 2022, 2024). We also contribute to the growing research interest in studying AI by emphasising the necessary work in contextualising data so that they are valuable to AI, as opposed to focusing on whether data, as AI outputs, are valuable to domain experts (Felin & Holweg, 2024; Lebovitz et al., 2021; van den Broek, 2025; van den Broek et al., 2021; Waardenburg & Sergeeva, 2021). For practice, making context explicit in platform schemas is necessary but not sufficient: what moves through must also support context reconstruction downstream.

Chapter 2

Research Background

This thesis sits at the intersection of two areas in information systems (IS) research: data studies and artificial intelligence (AI). The empirical focus is on how interpretive subsurface data is disambiguated through contextualisation in a Norwegian oil and gas operator, and how that work changes as the operator moves to the OSDU data platform. The literature reviewed here builds up the concepts the thesis uses to analyse that process.

The chapter has three sections. The first covers data: how IS research has conceptualised the term, and why the inherited view of data as raw facts about the world has been challenged (Aaltonen et al., 2023; Alaimo et al., 2020; Jones, 2019). The second covers data work: the practices through which data are produced and made meaningful, drawing in particular on studies grounded in offshore oil and gas (Mikalsen & Monteiro, 2021; Østerlie & Monteiro, 2020; Parmiggiani et al., 2022). The third covers debates on AI and data, focusing on how AI systems depend on data work and domain expertise to produce meaningful outputs, particularly in uncertain and knowledge-intensive domains (Felin & Holweg, 2024; Lebovitz et al., 2021; van den Broek, 2025; van den Broek et al., 2021; Waardenburg & Sergeeva, 2021).

This progression is analytically motivated. Understanding data conceptualisation provides the foundation for analysing what changes when interpretive data

is transformed and moves across disciplines and organisational boundaries. Examining data work reveals the practices through which disambiguation is achieved and maintained. Finally, considering how data relates to AI attends to some of the current challenges with disambiguating data in the industry. Together, these build toward the research question guiding this thesis: how industrial data is disambiguated through contextualisation, particularly as it departs from its origin and is reused in other settings (RQ).

An important boundary defines this review's scope. Although OSDU serves as useful empirical ground for our research question, we deliberately avoid engaging with platform– and ecosystem literature. The platform and its adoption is still maturing, and studying its platform dynamics would pull the thesis off course. We therefore refer to OSDU as it refers to itself – a data platform – but strictly in the colloquial sense. While it could certainly be an interesting perspective for this thesis, currently, that research stream falls outside our scope.

2.1 Data

What data refers to in IS research is less settled than the term's frequent use suggest. Jones (2019) observes that IS usage of the term has historically remained inconsistent. Recent workshops on the future of IS data research have flagged 'the semantic ambivalence of data in contemporary contexts' (Xu et al., 2025) and called on the field to 'go beyond traditional ways of conceiving data' (Aaltonen et al., 2023).

The 'traditional ways of conceiving data' that have historically anchored IS work on data treat data as 'raw' facts about the world: objective, neutral observations to be collected, stored, and processed into information and, eventually, knowledge (Jones, 2019). The Data–Information–Knowledge–Wisdom (DIKW) hierarchy gives this view its most familiar shape, placing data as the raw material at the base of a cognitive pyramid (Ackoff, 1989). This stance, described by Kitchin (2022,

p. 5) as a 'pre-factual' view on data is often contrasted with Gitelman (2013)'s formulation that 'raw' data is an oxymoron. Data are never simply found; they are always produced under conditions that shape what they can subsequently be. Tuomi (1999) argued two decades earlier that the conventional DIKW hierarchy should be reversed, because what counts as data depends on prior commitments about what is worth measuring.

The factual view fails most visibly in domains where the underlying phenomenon is not directly accessible and where data carry the residue of substantial prior interpretive work. Subsurface data, as shown in Chapter 1, is one such domain. There is no stable, observable referent against which a measurement can be checked. The hydrocarbons that the data are attempting to represent are unreachable kilometres under the seabed. In a strong sense, the data themselves are the phenomenon as it appears in everyday work (Mikalsen & Monteiro, 2021), and as a result, *how* we know the data becomes *what* we know (Monteiro, 2022, p. 21). As such, treating data as simply 'facts about the world' misses the important distinction that their meaning becomes dependent on the interpretive context they were produced in, and that context is not necessarily explicit in the data as we consume it.

A growing body of IS research takes this contingency as its starting point. Alaimo et al. (2020, p. 164) defines data as 'sign tokens used to describe, index, represent or stage (perform) reality', a definition that foregrounds the mediating role of data rather than treating data as a transparent window onto the world. Data are produced, and their value emerges from sociotechnical processes that aggregate and structure them for particular purposes (Aaltonen et al., 2021). This thesis is concerned specifically with digital data, and as digital artifacts they carry properties that shape how they circulate. Aaltonen et al. (2021) characterise digital data as editable, portable, and recontextualisable: data can be amended and recombined at low cost, can travel across settings and platforms, and can be put to uses other than those of their origin. The same properties that allow data to circulate also decouple them from the conditions under which they were produced,

making their meaning harder to fix as they move.

Two related framings of data's representational character are relevant for our empirical analysis. Representation Theory (Recker et al., 2019) holds that an IS's basic purpose is to represent real-world phenomena faithfully enough that users can reason about them without observing the phenomenon directly. Its representation model frames system development as a chain of transformations from human-oriented representations towards machine-oriented ones, emphasising representational fidelity as a requirement that has to be maintained across the chain. The semiotic taxonomy from Bailey et al. (2012), drawing on Peirce et al. (1931), classifies data as indices (direct correspondence with a referent), icons (resemblance), or symbols (relation by convention). Their case study demonstrates the effect of working with representations with varying levels of correspondence with the phenomenon it represents. Both framings matter for the empirical analysis. Subsurface data moves through long transformation chains, and at each step, the representation drifts further from the hydrocarbon formations that the data are taken to describe. Contextualisation work then becomes in many ways related to maintaining the representational link to the phenomenon across this transformation chain.

2.2 Data Work

The recognition that data are constructed rather than given shifts attention to the practices through which data are produced, maintained, and made meaningful. Data Work is the term some IS researchers have used to name this terrain, drawing on insights from Science and Technology Studies (STS) about the often invisible labour involved in crafting and making data work for downstream purposes (Parmiggiani et al., 2022). The practices are varied and include 'janitor work' such as cleaning and preparing data, annotating and labelling, reconciling discrepancies, and deciding what counts as an outlier (Østerlie & Monteiro, 2020; Parmiggiani et al., 2022). They tend to be treated as a preliminary to the more foreground

work of analysis, but as the literature on data work argues, this is already where much of the analytical and interpretive work happens (Parmiggiani et al., 2022, 2024).

Parmiggiani et al. (2022) make this argument most directly in their study of the 'backrooms of data science'. Following data managers and geoscientists in an oil and gas context, they show that data is discovered, prepared and consumed through extensive negotiations among heterogeneous actors with different stakes in what the data are taken to mean. The backroom work of assembling datasets, deciding what to include, and resolving conflicts between sources is where the conditions for downstream analysis are established. Decisions taken in the backroom shape what later analyses are built on. The work is largely invisible to the consumers of the resulting datasets, but it determines whether those datasets are interpretively trustworthy.

Østerlie and Monteiro (2020), working in the same broad empirical setting, develops a complementary account focused on representational becoming. Their three mechanisms (noise reducing, material tethering, triangulating) describe how digital representations of physical phenomena are made organisationally real through ongoing work. The work is continuous because the conditions that support representational faithfulness can change. Sand sensors produce false alarms that have to be filtered out as conditions in the well change, and predictive algorithms need recalibration against post-hoc physical inspection. A representation that held up under one set of operating conditions can stop holding up under another. The data work that maintains representations is therefore part of what the representations are.

Mikalsen and Monteiro (2021) extend this perspective. Their study of offshore explorationists working with seismic and well data identifies three patterns of work practice (accumulating, reframing, and prospecting) through which practitioners build provisional confidence in data whose relationship to the underlying phenomenon is fundamentally uncertain. The aim of the work is not to eliminate uncertainty but to make it actionable. Practitioners forge consistency

where consistency does not naturally exist, as in the well tie-in problem, where seismic data measured in time and well data measured in depth have to be reconciled despite a non-linear conversion. They fill in gaps using analogues from neighbouring formations, and reinterpret prior data when new evidence forces revision. What gives these practices their force is that they construct provisional grounds for action under conditions where resolution is not available.

Across these studies, the picture is consistent: data work in interpretive industrial domains is contextualisation work, producing data with their context and maintaining that context as the data move. Alaimo and Kallinikos (2022) argue that data objects gain value because they aggregate data and metadata into structured units that can be reused across contexts, presupposing that a well-structured data object will remain interpretively legible to consumers who were not party to its production. But reuseability requires that the relevant context be encoded in or alongside the data, which raises the prior question of which context counts as relevant, and for whom. Aaltonen et al. (2021) describes the work of answering that question as recontextualisation: fitting data-based objects to new interpretive settings through metadata, standards, and semi-automated processes. Parmiggiani et al. (2024) push this further temporally, arguing that data curation is anticipatory — practitioners encode metadata today on the assumption that future users will want to reconstruct who collected the data, when, and under what conditions, even though those users and purposes are not yet known. Recontextualisation, in this sense, is work done in advance against an interpretive context that does not yet exist, and its failures may not surface until much later.

2.3 Data and AI

AI is commonly framed as a data-driven approach to knowledge production, where patterns extracted from large datasets are expected to substitute domain expertise and human decision-making (Felin & Holweg, 2024; van den Broek, 2025; van den Broek et al., 2021). van den Broek et al. (2021) outlines how this paradigm

positions AI (and more specifically, machine learning) as capable of independently generating knowledge from data through the means of large-scale computation and complex pattern recognition, far exceeding human capabilities. This independent form of knowledge production further promises the benefit of mitigating human biases, cognitive limitations, and path-dependencies that have shaped experts' knowledge work (van den Broek et al., 2021). Machine learning is thus framed as a more objective alternative to expert judgment, grounded in the assumption that data is "raw" and reflects reality more accurately than human interpretation (van den Broek et al., 2021). In this sense, AI's mode of knowledge production assumes that data alone can function as a stable and sufficient basis for knowledge.

This form of knowledge production is not limited to supervised machine learning but extends to large language models (LLMs). However, as Felin and Holweg (2024) argue, these systems are limited by their probabilistic and "backwards-looking" nature, generating outputs by drawing on patterns in historical data rather than reasoning causally about why phenomena occur and which variables matter in a given situation. While such systems can appear intelligent through memorisation and their ability to express the same information in indefinite ways, their predictive and imitative nature is ill-suited for settings characterised by uncertainty and unpredictability (Felin & Holweg, 2024). These settings depend on forward-looking reasoning, in which knowledge cannot be derived from past data alone but rather through theories and experimentation (Felin & Holweg, 2024). Felin and Holweg (2024) challenges the assumption that data can serve as a stable and sufficient basis for knowledge, arguing that because it only "mirrors the past" and is inherently "theory-dependent", it cannot independently identify its own evidence or provide a stable ground truth in uncertain domains.

The assumption that data provides a stable foundation for AI is further challenged by Lebovitz et al. (2021), who highlights the inherent instability of "ground truth" labels used to train and validate these systems. Their study of different AI tools for medical diagnosis demonstrates that ground truth is not an objective reflection of reality, but a socially constructed artefact that does not capture the complexity of

professional work (Lebovitz et al., 2021). This instability stems from the inherent disconnect between "know-what" — the codified, explicit knowledge captured in labels — and "know-how" — the tacit, situated practices experts rely on to deal with uncertainty (Lebovitz et al., 2021). In practice, professionals often disagree on the "correct" label for identical inputs, revealing that knowledge is not fixed but contingent and subject to interpretation, consequently making the "ground-truth" labels subject to deep underlying ambiguity (Lebovitz et al., 2021). Because both data and expert knowledge are inherently unstable, Lebovitz et al. (2021) argue that AI tools need to be trained and validated on quality measures that more closely resemble real-world know-how practices. This highlights that data alone cannot serve as a fixed, neutral foundation that the AI paradigm tends to assume. The data itself and the expertise used to interpret it remain open to revision. AI systems therefore do not operate independently of human judgment but rely on continuous interpretive work to make both the data and their outputs meaningful in practice.

This is partly due to the "black box problem" of machine learning (ML), where how data is used and connected remains difficult for humans to discern (Waardenburg & Sergeeva, 2021). Waardenburg and Sergeeva (2021)'s study on using ML for crime detection highlights how "algorithmic brokers", tasked with the knowledge-intensive task of translating abstract predictions, enabled contextualised predictions that yielded meaningful information for end-users. Producing meaningful predictions required brokers to engage with both the machine learning community (developers) and the user community (police officers), drawing on knowledge from each to contextualise the outputs (Waardenburg & Sergeeva, 2021).

van den Broek (2025) further unpacks this knowledge work, showing that using AI at work requires multiple, interrelated forms of labour: data work, knowledge work, and values work. Building on the concept of data work as the invisible labour of cleaning and contextualising datasets (Parmiggiani et al., 2022), they show how AI initiatives intensify and redistribute this work, as the need for large volumes of

high-quality training data creates new pressures for datafication across domains previously untouched by such requirements (van den Broek, 2025). As shown in their study on using AI tools for hiring, this shift becomes an organising principle that reshapes professional roles, where practitioners take on ongoing responsibility for producing and structuring data that is suitable for AI (van den Broek et al., 2021). At the same time, this data work is increasingly intertwined with knowledge work and values work. As van den Broek (2025) and Waardenburg and Sergeeva (2021) show, AI systems depend on people to validate, adapt, and translate their outputs so that they become meaningful in practice, while also negotiating which values are embedded in these systems and which are left out. Decisions about what data to include and how to represent it, therefore, involve ongoing judgments about what counts as relevant information and how it should be interpreted (van den Broek, 2025). Consequently, rather than a linear process from data to knowledge, AI-based knowledge production emerges as an iterative and negotiated process shaped by ongoing work around data and its use in practice (van den Broek, 2025).

Summary

The three sections establish how data and knowledge are understood in IS research. Data are not raw facts about the world, but constructed artefacts whose meaning depends on the conditions under which they are produced. This construction is the result of ongoing work by practitioners, tools, and systems, where decisions about what to collect, how to represent it, and how to make it usable shape what data can subsequently become. In interpretive domains, this work is often complex and remains largely invisible, even though it plays a central role in making data meaningful in practice.

The literature on AI builds on and extends these insights. While AI is often framed as a data-driven approach to knowledge production, it relies on data that are themselves uncertain and subject to interpretation. Studies show that so-called

"ground-truth" is not stable, and that AI outputs do not speak for themselves but require interpretation and alignment with domain knowledge. Hence, AI does not eliminate the work involved in producing and making sense of data, but rather intensifies and redistributes it across different forms of labour.

A summary of the research background can be found in table 2.1. While these studies highlight the importance of the work through which data and knowledge are produced, we angle our study towards how such work supports making data interpretable and usable across different domains, systems, and over time.

Table 2.1: Summary of Research Background

Theme	**Related literature**	**Takeouts**	**Research gaps**
Data	Aaltonen, 2023; Aaltonen et al., 2021; Ackoff, 1989; Alaimo et al., 2020; Bailey et al., 2012; Gitelman et al., 2013; Kitchin, 2014; Recker et al., 2019; Jones, 2019; Tuomi, 1999; Xu et al., 2025;	• Data are situated and constructed. • Digital data can circulate while decoupling from its origins. • Fidelity costs are incurred as data move through transformation chains.	Which aspects of interpretive context can travel with data and which cannot.
Data work	Mikalsen et al., 2021; Parmiggiani et al., 2022; Alaimo & Kallinikos (2022); Parmiggiani et al., 2023; Østerlie & Monteiro, 2020;	• Data work is largely invisible and prior to the analysis it enables. • Context can be encoded in data in anticipation of unknown future reuse.	How context is maintained or lost across domains, particularly when original practitioners and tools are no longer available.
AI and domain knowledge	Felin & Holweg, 2024; Lebovitz et al., 2021; van den Broek, 2025; van den Broek et al., 2021; Waardenburg & Sergeeva, 2021;	• AI presupposes data as stable and sufficient for knowledge production. • Ground-truth labels are socially constructed • AI redistributes data work.	Attention has been centered mostly on AI outputs, rather than the data they consume.

Chapter 3

Methodology

This chapter outlines the methodological approach used in the study. Section 3.1 presents the research design and justifies the case selection. Section 3.2 describes the process of data collection, including how empirical material was generated. Section 3.3 explains the data analysis, focusing on how the material was interpreted and structured to address the research question. Here, we also present our Gioia-inspired data structure, as seen in Figures 3.3 and 3.4, which provides the basis for the presentation of the findings in Chapter 5. Finally, Section 3.4 assesses the research quality in terms of validity, reliability, and ethics. Figure 3.1 provides an overview of the key steps in the research process and how they relate to each other.

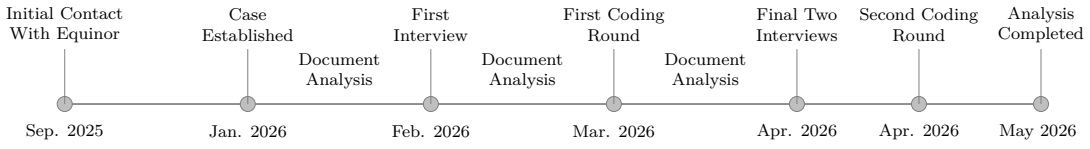

Figure 3.1: Timeline of key activities of this study's data collection and analysis

3.1 Research Design & Case Selection

Our methodology follows a qualitative single-case study design (Gerring, 2004). The case centres on Equinor, a Norwegian oil and gas operator. In this thesis, we examine how subsurface data are disambiguated through contextualisation across the subsurface workflow, and how these processes are reorganised with the introduction of a new standardised data platform and AI-related initiatives. Equinor represents a relevant case due to its active involvement in the digitalisation of subsurface work, including its engagement with the Open Subsurface Data Universe (OSDU) initiative. At the same time, the case reflects broader industry developments related to the standardisation of subsurface data and the growing integration of AI into established workflows. The case is therefore well-suited for examining how industrial data are disambiguated through contextualisation as they depart from their original settings and become reused across workflows and domains.

A qualitative single-case study is appropriate because the study seeks an in-depth understanding of how contextualisation processes unfold within a real-world organisational setting (Gerring, 2004). Following Gerring (2004), the case study allows us to examine how subsurface data is contextualised and disambiguated across workflow stages and into a shared data platform. The study adopts a process perspective by following subsurface data from data acquisition to business decision-making, enabling an analysis of how interpretive conditions are shaped and a re-established across sequential stages of the workflow. In addition, the study incorporates a temporal dimension by examining how this work is reconfigured in the transition from application-bound data environments to OSDU. Access to both internal subsurface workflows and ongoing platform implementation efforts makes the case particularly valuable for studying how subsurface data is interpreted and reused across contexts.

Finally, we adopt an exploratory approach grounded in an interpretive epistemology, as we seek to understand how subsurface data are disambiguated

within organisational practice. Interpretive research in IS emphasises how meaning is shaped within specific organisational contexts, where information systems are embedded in ongoing practices and social processes (Myers, 1997). This is particularly relevant in our study, where contextualisation is shaped both by subsurface workflows and by the transition toward the OSDU data platform. As the role of contextualisation in the disambiguation of industrial data remains underexplored in IS research, the study follows an open and iterative approach in which concepts are developed through engagement with the empirical material rather than defined in advance. In line with this perspective, interviews and documents are treated as interpretive sources through which participants express how they understand and work with subsurface data across the workflow. We position the study within internal realism (Walsham, 1995), assuming that the subsurface exists materially, but that access to it is mediated through actors' interpretations and representations.

3.1.1 Case Access and Prior Engagement

Case access to Equinor was established through the course IN5560 - Data Governance at the University of Oslo, where both authors collaborated with the company while writing a project report analysing its data governance approach. Through this collaboration, we engaged with a contact person at Equinor who holds roles as Lead Digital Architect within the organisation and as a board member of the OSDU Forum. The contact person facilitated access to the organisation and assisted in identifying relevant informants with expertise in data management and data governance for the project report, and within data management, data science, IT, geology, and petrophysics for the master's thesis.

As part of the coursework, five interviews were conducted with Equinor employees on topics related to data governance. While these interviews provided valuable background understanding of the organisation and its data landscape, they do not form part of the empirical material analysed in this thesis. The primary empirical

data consist exclusively of the interviews conducted specifically for this study, which are described in detail in the following section.

3.2 Data Collection

The data collection consisted of two data sources: semi-structured interviews and document analysis. In total, we conducted nine semi-structured interviews. The documents analysed include company presentations, industry and regulatory reports, and OSDU documentation. Semi-structured interviews served as the primary source of empirical material, while documents were used to contextualise the case organisation, its subsurface workflows, and the OSDU initiative. The data was collected between January 2026 and April 2026, where data collection overlapped with data analysis.

3.2.1 Data Collection Stages

The insights from our prior engagement, as described in Section 3.1, provided valuable contextual understanding of the case organisation and informed the design of the master's thesis study. In collaboration with the contact person at Equinor, we discussed which aspects of the organisation could be interesting to investigate within the time frame of the thesis. These discussions resulted in a focus on the subsurface workflow, AI deployment, and the OSDU initiative.

Following these discussions, the contact person helped us identify relevant informants within the subsurface domain and related data initiatives. In this phase, we conducted five exploratory semi-structured interviews, which helped shape the research focus. Here, we aimed at understanding subsurface workflows, how and why OSDU came to be, and the current and potential role of AI in the subsurface domain. After this initial exploratory phase, we conducted two additional interviews with a narrower scope. These later interviews focused on

topics such as the meaning of context in the subsurface domain, how OSDU supports generative AI, and the mechanisms through which OSDU preserves contextual information, helping clarify emerging themes and refine the direction of the study.

Following the first seven interviews, we began analysing the empirical material that had been collected. These early insights informed the direction of the final two interviews, where the preliminary analysis allowed us to narrow down the research focus further. In these interviews, we explored topics that had emerged as particularly relevant during the analysis, including the loss of context across workflow stages, trustworthiness of data, and how OSDU is used presently to explicate context.

3.2.2 Semi-structured Interviews

The main source of empirical data was semi-structured interviews. In total, nine interviews were conducted, each lasting between 45 and 75 minutes. All interviews were held digitally over Teams, due to geographical constraints. Access to interview participants was facilitated through the study's contact person at Equinor. Through this collaboration, we were able to identify relevant informants with expertise in petrophysics, geology, data management, data science, and the OSDU data platform. An overview of the informants is provided in Table 3.1. The interviews were semi-structured to balance structure and flexibility. This format allowed informants to reflect on their experiences and perspectives while ensuring that the discussion remained aligned with the research objectives. The interview guide, therefore, was organised around thematic topics, each supported by predefined questions, while still allowing for follow-up questions and exploration of topics that emerged during the conversation. An example interview guide is included in Appendix A.

All interviews were conducted jointly by both authors, Eirik and Stefan.

Table 3.1: Overview of interviews and informants

Informant	Role	Phase	Main focus
A	Petrophysicist	Exploratory	Static reservoir modeling, subsurface workflows, context loss, and AI potential
B	Geologist	Exploratory	Subsurface workflows, sources of uncertainty, and AI potential
C	Data Engineer	Exploratory	Data engineering and governance
D	Data Analyst	Exploratory	The technical implementation of OSDU and why it came to be
E	Data Scientist	In-depth	Generative AI and its use cases within the subsurface domain, how OSDU enables Generative AI
D	Data Analyst	In-depth	OSDU capabilities and mechanisms for preserving context
F	Data Manager	In-depth	Meaning of context within the subsurface and the role of OSDU in explicating this context
A	Petrophysicist	In-depth	Loss of context across workflows, uncertainty handling, and how OSDU attempts to explicate context
E	Data Scientist	In-depth	Foundation models, GenAI as a means to becoming more efficient, and how OSDU is used today

Conducting the interviews together allowed one author to lead the conversation while the other took notes and asked follow-up questions. We also debriefed immediately afterwards to identify relevant topics and clarify observations. This division of roles made it easier to remain attentive to the informants' perspectives and facilitated shared reflection on the interview material. In addition to being documented through notes, the interviews were also recorded using Microsoft Teams. All informants were informed about the recording of the interviews and provided consent through a written consent form (see Appendix B). The recordings were transcribed using the University of Oslo's 'autotekst' transcription service, which enabled systematic analysis of the interview material and supported the subsequent coding process.

3.2.3 Document Analysis

In addition to semi-structured interviews, we used document analysis as a secondary data source. The documents included company presentations, industry and regulatory reports, and OSDU documentation. The document analysis provided a systematic approach to reviewing and interpreting documents and supported triangulation with other empirical material (Bowen, 2009).

Document analysis is advantageous in qualitative research because documents are non-reactive sources of data, meaning that they are produced independently of the research process and are therefore not influenced by the presence of the researcher (Bowen, 2009). However, documents may be incomplete, outdated, or reflect official narratives rather than actual practices within the organisation (Bowen, 2009). For this reason, the document analysis was used primarily to complement and contextualise the primary empirical material obtained through interviews.

Industry and regulatory reports provided insight into the broader regulatory and technological developments related to digitalisation, data management, and AI in the sector. While company presentations contributed to understanding subsurface workflows and related challenges. In addition, OSDU documentation provided insight into the platform's architecture, data models, and intended use within the subsurface domain. The documents also helped familiarise us with concepts and terminology relevant to the study, such as technical assurance and activity templates.

Both authors reviewed the documents independently before discussing them jointly to align interpretations and identify relevant themes and contextual insights. The documents were also used to triangulate interview-based interpretations by comparing official descriptions of workflows, systems, and platform features with how participants described their practical experiences and everyday work.

3.3 Data Analysis

This section outlines the analytical approach used in the study. The analysis began inductively, drawing on first-order categories (Gioia et al., 2013) and a categorising strategy (Maxwell & Miller, 2008) to stay close to the informants' perspectives. As the analysis progressed, we adopted an abductive and iterative approach, moving between empirical data and relevant literature (Dubois & Gadde, 2002). Drawing on Maxwell & Miller's connecting strategy (Maxwell & Miller, 2008), we examined how categories related to one another across interviews and stages in the subsurface workflow, introducing a process-oriented perspective that traces both the movement of data across workflow stages and the ongoing transition toward a shared data platform. This involved repeatedly revisiting earlier coding as new interviews were conducted, refining first-order concepts, and reassessing second-order themes and aggregate dimensions in light of new insights (Gioia et al., 2013; Maxwell & Miller, 2008).

This resulted in a data structure inspired by Gioia et al. (2013), which supported the organisation and abstraction of empirical material into concepts and themes, while the connecting strategy guided the interpretation of relationships across workflow stages and the transition from application-bound systems toward OSDU. Together, these analytical steps contributed to the development of a process model capturing how industrial data are disambiguated through contextualisation as they move from local source systems toward a shared data platform and are reused in other settings.

3.3.1 Analytical Process

In order to structure the analysis in the initial inductive stage, we applied selected elements from the Gioia methodology (Gioia et al., 2013). Rather than adopting the methodology in its entirety, this study utilises its elements and applies the Gioia data structure as an analytical device to organise and present empirical findings

through first-order concepts, second-order themes, and aggregate dimensions.

To complement the structuring logic provided by the Gioia methodology, our analysis draws on categorising and connecting strategies (Maxwell & Miller, 2008). The categorising strategy supported the process of grouping segments of data based on similarity, through coding and the development of themes (Maxwell & Miller, 2008). This aligns closely with the use of first-order concepts and second-order themes in the Gioia data structure.

Connecting, in contrast, focuses on relationships between data segments. Rather than separating data solely into categories, we sought to preserve how elements are related within a broader context (Maxwell & Miller, 2008). Maxwell and Miller (2008) emphasises that qualitative analysis should not rely only on categorisation, as this can miss important contextual relationships in the data.

These strategies were used in combination throughout the analysis. The Gioia data structure and the categorising strategy supported the abstraction and organisation of empirical material, while the connecting strategy introduced a process-oriented perspective focused on how contextualisation processes evolved as data moved across the subsurface workflow and into the shared data platform.

First-order Categorisation

Initially, we individually read the interview transcripts multiple times as the interviews were conducted to study and familiarise ourselves with the empirical material. This was supplemented with analytical memos, summaries shared with informants to avoid misunderstandings, and a review of relevant documents. The purpose was to understand the terminology used by the informants and identify initial patterns in the data.

Following the first seven interviews, we began inductive first-order coding of the interview transcripts. Both authors independently identified and coded empirically relevant data segments related to subsurface workflows, data

movement, interpretation, and uncertainty. At this stage, the aim was to remain close to the empirical material and avoid abstracting too early.

In parallel, we introduced a simple form of connecting analysis through colour-coding. Rather than functioning as fixed codes, the colour categories acted as analytical lenses that helped trace how topics such as uncertainty, contextualisation, and interpretation emerged across workflow stages and informants. This supported the identification of relationships based on contiguity (Maxwell & Miller, 2008), for instance, how uncertainty and loss of interpretive context emerged in relation to specific data transformations or handovers.

After this initial coding phase, the material was structured and compared across interviews. We discussed and reconciled individual codes collaboratively, revising them until a shared understanding of the empirical material was reached. Throughout the process, we aimed to keep the codes close to the informants' own language and interpretations. This resulted in 205 informant-centric first-order categories.

Development of Second-Order Themes through Categorisation

We continued with a categorising strategy by clustering first-order concepts into more conceptual second-order themes (Gioia et al., 2013). Following Gioia et al. (2013), we increasingly adopted the role of "knowledgeable agents" by interpreting patterns across informant-centric codes. The clustering focused on recurring challenges related to how subsurface data were interpreted and disambiguated, as well as how these challenges were related to the transition toward OSDU. This was done along three main dimensions.

First, we examined whether concepts described similar underlying challenges related to interpreting and working with subsurface data across the workflow. This included issues such as the loss of embedded context when data moved out of proprietary applications, the compounding uncertainty across modelling stages, and the need to reconcile heterogeneous data types. We examined how these

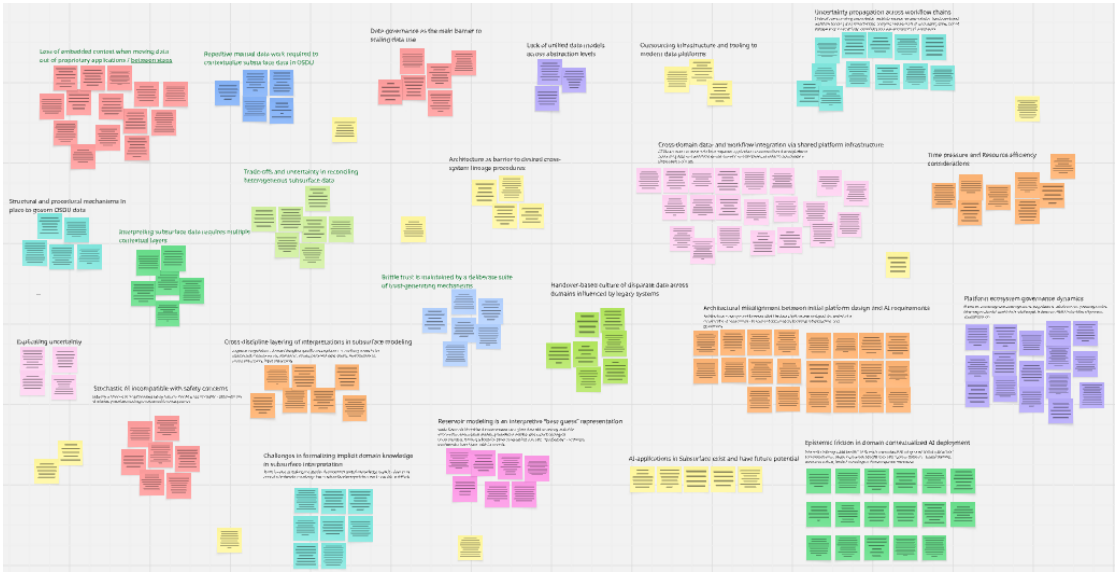

Figure 3.2: Screenshot of our in-progress clustering of codes using a digital whiteboard application.

challenges became more pronounced in relation to the transition toward OSDU, where data became increasingly detached from the domain-specific applications and practices that originally anchored their meaning, creating new challenges for maintaining meaning and trust when contextual information was no longer embedded in specialised tools.

Second, we compared how informants described dealing with these situations, grouping concepts that reflected similar approaches of preserving, reconstructing, or making context explicit. This included standardising and structuring data models, stabilising and reconstructing contextual relationships, maintaining trust in the data, and connecting datasets to re-establish meaning across domains. These processes were often closely tied to the introduction of the shared data platform, where contextual information had to be externalised and standardised rather than remain embedded within applications and individual expertise.

Third, we considered where in the subsurface workflow the concepts occurred, which allowed us to identify recurring patterns across stages and transitions. This included how interpretive representations are produced at multiple stages of the workflows, how models were constructed as “best-guess” representations under uncertainty, and how handover-based work practices contributed to fragmentation

across stages of the workflow. This comparison made visible how disambiguation increasingly depended on reconstructing contextual relationships as data moved from application-bound systems toward cross-domain forms of reuse, including AI-related use cases.

This initial clustering resulted in 23 second-order themes. As the final two interviews were conducted, we continued the categorising strategy while iteratively refining the set of categories by merging, splitting, and reinterpreting them in light of new insights. In this process, we assessed the extent to which each category contributed to understanding how subsurface data were disambiguated through contextualisation as they moved across workflow stages and toward a shared data platform. Through this iterative refinement, we chose to proceed with 13 second-order themes. The remaining categories were not discarded as unimportant, but set aside as they primarily captured organisational and infrastructural aspects surrounding the OSDU platform rather than the interpretive and contextual processes at the core of this study.

Abductive Iteration and Connecting into Aggregate Dimensions

At this stage, the analysis became more explicitly abductive (Dubois & Gadde, 2002) as we moved iteratively between the empirical material and relevant literature on data, uncertainty, digital representations, data work, and AI in IS research. While the second-order themes captured recurring empirical patterns, they did not fully explain how these patterns became interconnected across the subsurface workflow.

The literature sensitised us to how data become meaningful through situated interpretation and ongoing data work. Prior studies highlighted that data are situated and constructed and that contextual information can be encoded in anticipation of future reuse. However, the literature provided less insight into how interpretive context is maintained or lost as data depart from their original settings and become reused in other settings. This became increasingly relevant in

our empirical material, where interviews described challenges related to preserving interpretability as subsurface data moved across applications and workflow stages.

Returning to the empirical material with this in mind, we observed that contextual information was often embedded within applications, workflows, and expert knowledge, but became fragmented as data moved across workflow stages. This became particularly visible in relation to the transition toward OSDU, where assumptions and relationships that were previously embedded within local tools increasingly had to be reconstructed through shared data models and platform features. In this setting, disambiguation depended on making contextual relationships explicit so that data could remain interpretable and reusable across settings.

To further examine these relationships, we applied what Maxwell and Miller (2008) refer to as a connecting strategy. Rather than grouping themes based only on similarity, this step focused on identifying relationships and sequences between themes (Maxwell & Miller, 2008). We conducted a visual mapping exercise using a whiteboard, where we iteratively placed the themes along the subsurface workflow and within the broader temporal development of the case, including the ongoing transition toward the OSDU data platform. In this process, we focused specifically on how themes related to (1) how interpretive context is embedded within domain-specific applications and workflows, (2) the detaching and fragmentation of this context as data is moved between applications and stages of the workflow, and (3) the re-establishment of context within the shared data platform.

This analysis revealed that several second-order themes were interconnected through shared dynamics rather than representing isolated phenomena. Themes related to data handovers, application-bound representations, and loss of embedded context were connected through a broader dynamic of fragmentation across workflow stages. Similarly, themes related to reconstructing and stabilising interpretive context in OSDU and the ongoing work required to maintain data interpretability were connected through a broader process of reconstructing context across domains within the shared data platform. Other connections

followed similar logic, where themes became linked through broader processes of contextualisation and reuse across the subsurface workflow.

Through the connecting strategy, the analysis moved from a set of categorised themes to a more integrated, process-oriented understanding of the case. Rather than abstracting themes independently, the aggregate dimensions emerged from the relationships identified between themes across workflow stages and the ongoing transition toward OSDU, resulting in five aggregate dimensions (Gioia et al., 2013). Two dimensions captured context as embedded within proprietary applications and workflows, while three dimensions captured how context increasingly becomes reconstructed and maintained through the shared data platform.

Importantly, these dimensions should not be understood as representing a clear shift from one mode of contextualisation to another. Rather, they reflect an ongoing transition in which established domain-based forms of interpretation continue to coexist with emerging forms of platform-based contextualisation. The aggregate dimensions, therefore, synthesise both the continuity of existing interpretive work and the changing processes through which data are disambiguated as they become increasingly detached from their original environments and reused across settings.

3.3.2 Data Structure

Figures 3.3 and 3.4 present the data structure developed through the analytical process.

In this study, the data structure reflects both the categorising and connecting stages of the analysis. The left column shows the informant-centric first-order categories, which capture how participants described their experiences in their own terms. These were grouped into second-order themes in the middle column, where patterns across interviews were interpreted and abstracted.

Building on this, the connecting analysis examined how the second-order themes

were related across the subsurface workflow and the broader transition toward the OSDU data platform. This resulted in the aggregate dimensions shown in the right column. These dimensions represent the highest level of abstraction in the analysis, capturing broader processes related to contextualisation, interpretability, fragmentation, reconstruction, and trust in subsurface data.

Overall, the data structure serves as a visual representation of the links between data and interpretation, making explicit how informant-centric categories were progressively abstracted into a process-oriented analytical framing. It thereby enhances transparency and qualitative rigour by allowing the reader to trace the connections between raw data and the study's findings, and serves as the foundation for the presentation of the findings in Chapter 5.

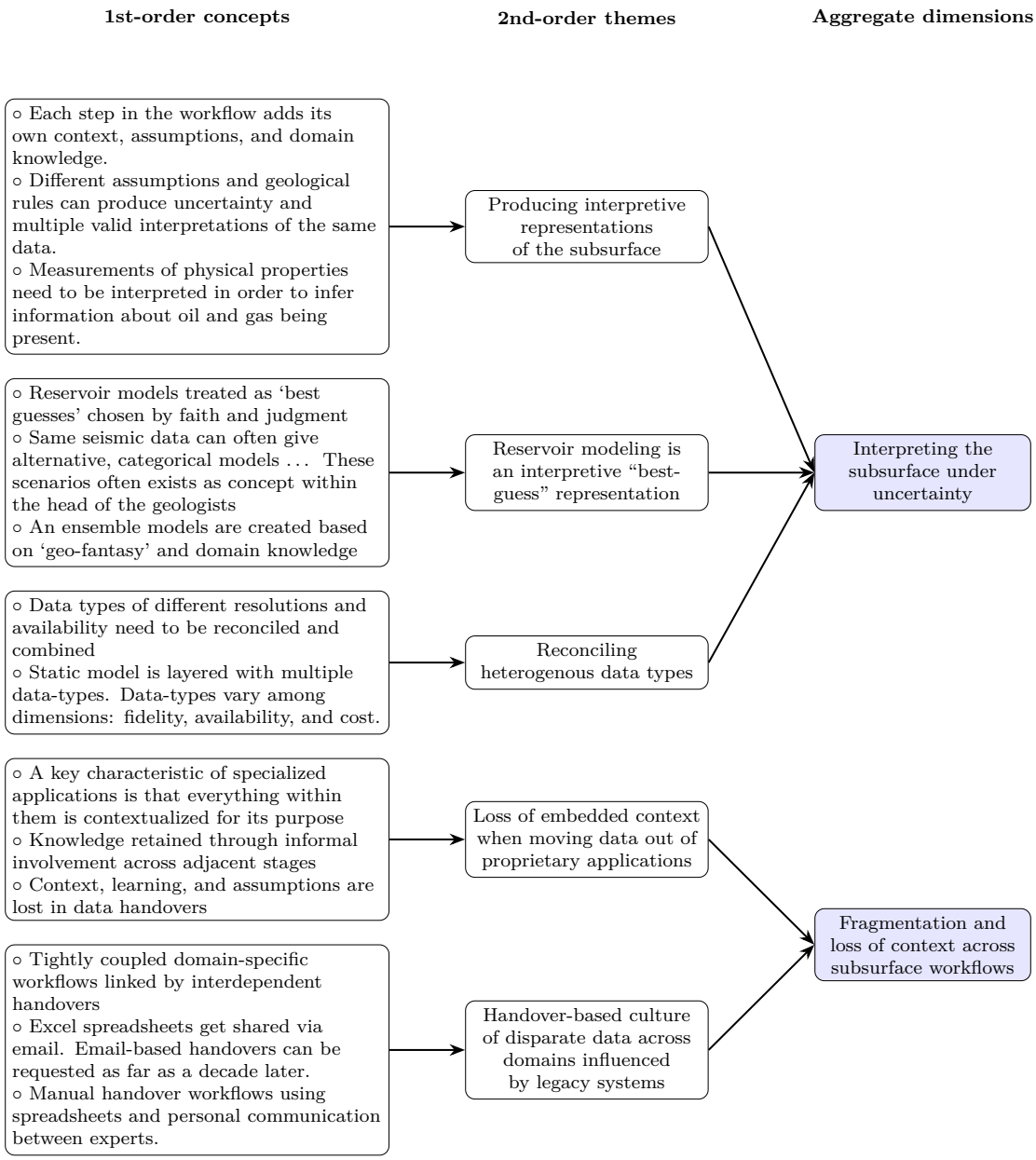

Figure 3.3: First part of the data structure

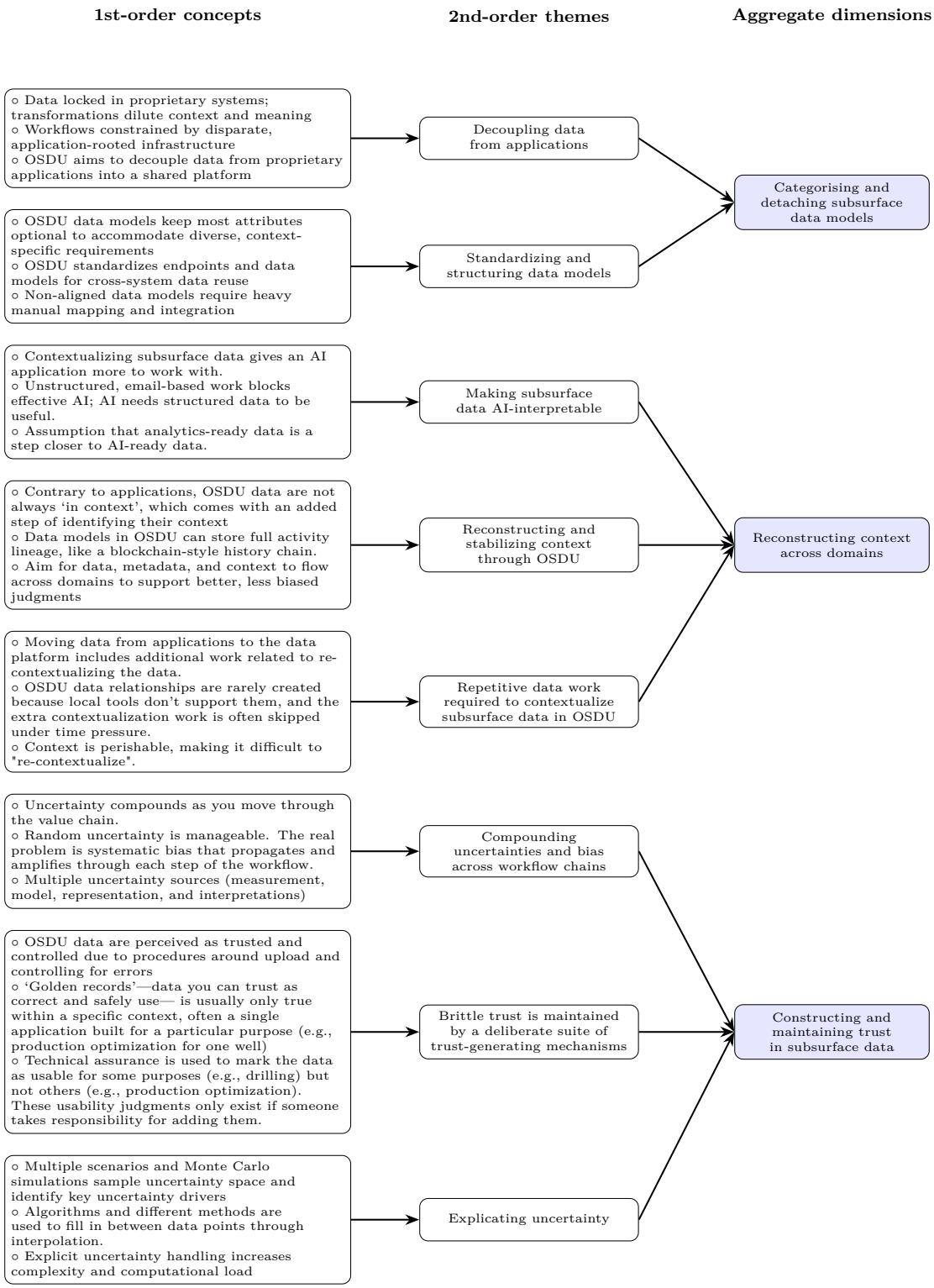

Figure 3.4: Second part of the data structure

3.3.3 Data Structure to Process Model

Marking the culmination of our data analysis, we used the Gioia-inspired data structure as a basis for developing a process model capturing the broader dynamics and interrelationships identified across the empirical material. While the data structure visualises the progression from first-order categories to second-order themes and aggregate dimensions, the process model captures how subsurface data are disambiguated through contextualisation as they move away from their original settings and become reused in other settings. In particular, the model captures how contextual information becomes categorised, detached from their original environments, and subsequently reconstructed across domains in order to maintain interpretability, trust and support reuse. This resulted in Figure 6.1, which constitutes the main contribution of the study, and is elaborated on further in the discussion in Chapter 6.

3.4 Research Quality

The case study method is sometimes criticised for lacking rigour in how findings are derived (Gibbert et al., 2008). This makes it important to address such concerns in our study. To do so, we draw on Yin (2018), who proposes a set of tests for assessing case study quality: construct validity, internal validity, external validity, and reliability. As outlined in section 3.1, we consider our study to be exploratory rather than explanatory. Hence, internal validity is not considered an appropriate test (Yin, 2018).

Construct validity refers to how well a study defines and measures the concept it is trying to investigate. It ensures that the collected data actually reflects the phenomenon being studied, rather than the subjective interpretation of the researchers (Yin, 2018). Our study has ensured this in several ways. First, our analysis has been grounded in multiple sources of data, including interviews with domain experts and relevant industry documents provided by the informants. This

has enabled triangulation of the empirical data, which is an important measure to enhance construct validity (Gibbert et al., 2008). Second, in line with Gioia et al., 2013, we relied on informant-centric first-order concepts, which closely reflect how participants described their experiences. This reduced the risk of forcing the data into predefined categories and helped ensure that our concepts were closely grounded in what we observed. We also used a clear data structure to show how we moved from raw data to aggregate dimensions, creating a chain of evidence that makes it possible to trace our interpretations back to the data (Gibbert et al., 2008).

External validity is about the extent to which the findings are generalisable and can be applied to other settings beyond the specific case under study (Gibbert et al., 2008). In case study research, this is achieved through analytical rather than statistical generalisation, where findings are related to broader theoretical concepts (Yin, 2018). Rather than claiming that results apply broadly across settings, the aim is to contribute to a theory that may be applicable in similar contexts. Following Yin (2018), external validity in case studies is strengthened by linking empirical findings to theory during the research design and analysis process. In this study, analytical generalisation was supported through the iterative movement between the empirical material and relevant literature following the initial coding process. This enabled emerging findings to be interpreted in relation to broader theoretical discussions rather than remaining specific to the immediate case context. In addition, the study provides a detailed account of the subsurface workflow and the transition to OSDU, enabling readers to assess the relevance of the findings for similar industrial settings.

Reliability in case study research refers to ensuring that the same findings can be reached if a later study follows the same procedures, and to reducing the risk of random errors (Gibbert et al., 2008; Yin, 2018). Gibbert et al., 2008 emphasises transparency and replication as key mechanisms for ensuring reliability of case study research. We have aimed to achieve this by providing a clear and detailed account of the research process. This includes the research design and

case selection, the data collection procedures, and the analytical approach. The data analysis followed a systematic and transparent procedure where we combined elements from the Gioia methodology with categorising and connecting strategies. The analytical process is described step by step, including how codes, themes, and aggregate dimensions were developed and refined. Finally, the use of a data structure further strengthens reliability by providing a visual representation of how the findings were derived from the empirical material. As this is an interpretive study, where we treat ourselves as "knowledgeable agents" (Gioia et al., 2013), the data structure also serves to make the link between data and interpretation explicit. Together, these elements enhance the reliability of the study by ensuring transparency and traceability in the research process.

Ethics. This study was conducted in accordance with the research ethics guidelines of the University of Oslo. The project was registered with Sikt (Norwegian Agency for Shared Services in Education and Research), where we provided documentation of the research design, purpose of the study, data collection procedures, and how data would be stored and processed.

All participants received a consent form before the interviews (see Appendix A), outlining the purpose of the study, how their data would be used, and their rights as participants. Participation was voluntary, and informants were informed that they could withdraw from the study at any time without consequences. To ensure confidentiality, all data was handled and stored securely using OneDrive with our university email, and sensitive data about the informants was removed or anonymised in the reporting of the findings. Only the researchers and the supervisor had access to the raw data. These measures were taken to protect participants and ensure the responsible handling of sensitive information.

Chapter 4

Case Description

This chapter provides the contextual background for our study. It begins by outlining the industry setting in which the case is situated, before introducing the case company and its role in upstream oil and gas activities on the Norwegian Continental Shelf (NCS). Building on this, the chapter describes the subsurface workflow as the central unit of analysis, followed by a discussion of recent developments related to increasing data scale and efficiency demands. The chapter concludes by presenting the industry response to these developments through the OSDU initiative.

4.1 Industry Context

The oil and gas industry is a major contributor to the global energy market, where it acts as a primary fuel source for the global economy (Burclaff, 2025). The industry is commonly divided into three segments: upstream, midstream, and downstream. For this thesis, we will focus on the upstream segment in the Norwegian offshore sector, consisting of oil and gas exploration and production (Burclaff, 2025). Offshore oil and gas exploration activities on the NCS consist of conducting seismic surveys to map subsurface geology and acquiring license rights, followed by production activities involving offshore drilling. The processes

and systems involved in these activities are highly complex and capital-intensive, where the cost of seismic studies and drilling a dry well can range from 5 to 20 million USD or more per exploration site (Burclaff, 2025).

The Norwegian upstream oil and gas ecosystem spans multiple actors, including operators, the Norwegian State, cloud providers, software vendors, and regulatory bodies. Operators are responsible for exploration and production activities, while the state plays a central role through ownership structures and regulatory oversight. Technology providers, including cloud infrastructure providers and software vendors, support daily operations and subsurface workflows. These actors deliver the technical infrastructure and tools used across different stages of exploration and production.

4.1.1 Regulatory Framework

Norway's petroleum regime is organised around a licensing system in which the Ministry of Energy awards exploration and production licenses to groups of companies (NorskPetroleum, n.d.). These licensees form joint ventures, with one company appointed as the operator (NorskPetroleum, 2025). The operator is responsible for operational activities as outlined by the license, production reporting in compliance with regulations set by The Norwegian Offshore Directorate (Bøe, 2015) and health, safety, and environment requirements established by The Norwegian Ocean Industry Authority (Havindustritilsynet, 2024a).

In addition to this licensing structure, Norway maintains a state participation model through the State's Direct Financial Interest, where the Norwegian State holds ownership shares in selected production licenses. Through this model, the state receives a direct share of revenues and costs from petroleum production (NorskPetroleum, 2026).

The licensing system and state participation model are anchored in the Petroleum

Act, which also requires operators to submit raw seismic, well, and production data, including interpreted data, to the national DISKOS database (Bøe, 2015; Monteiro, 2022). The Norwegian Offshore Directorate also enforces an embargo regime in which raw data remain confidential for two years, while interpreted data remain confidential for five years (Sokkeldirektoratet, 2024).

In addition to these requirements, operators within the DISKOS consortium also share and exchange data beyond what is formally mandated (Monteiro, 2022). Data ownership is tied to the licensing system, where raw data generated under a production license is owned collectively by the licensees. Each partner holds a defined stake, and all partners have equal access rights to the data, while sharing with third parties requires agreement among the partners. The Norwegian Offshore Directorate further provides industry-specific reporting guidelines for well, geophysical, and production data, commonly referred to as the "blue", "yellow", and "green" books (Sokkeldirektoratet, n.d.).

4.1.2 Safety in Petroleum Operations

Offshore oil and gas exploration and extraction on the NCS are inherently high-risk activities. Accidents such as the Alexander Kielland disaster illustrate the potentially severe consequences of operational failures (Smith-Solbakken & Dahle, 2025). As a result, safety has become a central concern in petroleum operations, and Norwegian authorities have set an ambition for petroleum activities to be world-leading in terms of safety standards (Havindustritilsynet, 2024b). As mentioned, the Norwegian Ocean Industry Authority is responsible for overseeing health, safety, and working environment conditions in petroleum activities. This includes developing regulations, supervising companies, and acting as a competence body for industry and government (Havindustritilsynet, 2024b). Through these activities, operators are required to manage risk in a systematic manner across all stages of petroleum operations.

The mandate of ensuring safe and efficient operations increasingly extends to

digital technologies, including artificial intelligence (AI). The Norwegian Ocean Industry Authority has provided a knowledge overview of risks related to the development and application of AI in the oil and gas industry, with particular attention to major accident risk. The report highlights challenges related to data quality, explainability, model degradation over time, and risks associated with human–computer interaction (Markussen et al., 2024). These challenges are relevant across petroleum workflows, including subsurface activities such as seismic interpretation, reservoir modelling, and drilling-related decision-making. The adoption of AI in the petroleum sector, therefore, takes place within an environment characterised by regulatory requirements, safety considerations, and established risk management practices.

4.2 Case Company: Equinor

Equinor is an oil and gas operator specialising in upstream oil and gas operations on the NCS. The company has approximately 20,000 employees and operates in more than 20 countries, with over 50 years of experience in exploration and production. Equinor is responsible for approximately 70% of total oil and gas production on the NCS (Tollaksen et al., 2025).

Within this context, Equinor operates as a central actor in exploration and production activities. The company is involved in a wide range of activities related to exploration, field development, and production. This includes conducting seismic surveys, participating in licensing rounds, and carrying out drilling operations across multiple fields. Through its role as operator, Equinor is responsible for managing daily operations and coordinating activities across disciplines and partners within production licenses. These activities are carried out in collaboration with other license partners, where responsibilities and data are shared within joint ventures in accordance with the licensing framework described in section 4.1.

A key aspect of these operations is the role of subsurface work. Subsurface disciplines are central to value creation at Equinor, where reservoir understanding plays a critical role in shaping field development decisions, well placement, and recovery strategies.

The processes involved in oil and gas exploration and extraction at Equinor are highly data-intensive. Large volumes of subsurface data are handled in daily operations, where data can be generated at rates equivalent to thousands of movies per second (Equinor, n.d.-a). In practice, this involves activities such as interpreting seismic data, analysing well logs, and constructing reservoir models that support decision-making. At the same time, Equinor is in a position where historical subsurface data, stored in the national DISKOS database, continues to generate value, as previously collected data may still provide useful insights into geological formations and reservoir conditions. Equinor possesses more than 50 petabytes of subsurface data accumulated over several decades of exploration and production activity (Equinor, n.d.-b).

To support these activities, Equinor is actively engaged in digitalisation initiatives through the use of advanced subsurface tools and cloud-based infrastructure. The company is also a member of the industry-wide standardisation initiative, the Open Subsurface Data Universe (OSDU), where it contributes to architectural discussions and the development of data standards. As part of this work, Equinor has been involved in defining data schemas and developing models for seismic and well data to support interoperability across disciplines and software systems within the subsurface domain. This is elaborated on in section 4.4.

As a large upstream operator with a long operational history and an extensive subsurface portfolio, the company manages substantial volumes of both legacy and newly acquired subsurface data. These characteristics, in addition to their transition to the industry-wide standardisation initiative of OSDU, make Equinor a relevant case for this study.

4.3 The Subsurface Workflow

The subsurface workflow in oil and gas exploration can be understood as a sequence of activities that moves from data acquisition to business decision-making. A simplified representation of this workflow includes stages such as data acquisition, interpretation and modelling, production forecasting, and finally investment decisions.

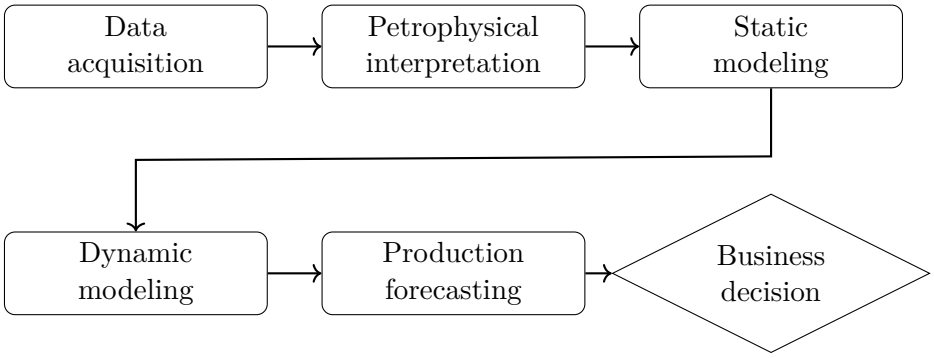

Figure 4.1: Simplified subsurface workflow from data acquisition to business decision-making.

The workflow typically begins with surveying the seabed to identify geological structures that may indicate the presence of hydrocarbons. Once a potential reservoir, or "prospect", is identified, exploration wells are drilled to determine whether hydrocarbons are present. If a discovery is confirmed, additional delineation wells are drilled to define the size, shape, and commercial viability of the reservoir (IADC, 2012). Based on this understanding, both static and dynamic reservoir models are constructed and used to evaluate different production scenarios. This forms the basis for production forecasting and economic evaluation, which ultimately informs investment decisions related to field development.

More specifically, the workflow involves data moving across specialised applications and disciplines, where the outputs from one stage become inputs to the next. As illustrated in Figure 4.1, vendor-acquired data form the starting point, before being

progressively interpreted, modelled, and aggregated into forecasts that support decision-making. Each stage produces new models, interpretations, and datasets that support subsequent stages in the workflow.

4.3.1 Subsurface Data and Disciplines

Seismic data are acquired through surveys of the seabed and consist of measurements recorded in time. These data cover large spatial areas and are used to construct a coarse representation of subsurface structures. However, they provide limited detail about specific reservoir properties. Seismic data are also highly processed, with only a small portion of the original signal retained after filtering and interpretation (Monteiro, 2022).

In contrast, well data are collected through drilling activities and provide detailed, localised information about the subsurface. These include measurements from sensors as well-logs or physical core samples of rock formations. Well data are often considered the closest available representation of subsurface conditions, but they are sparse due to the high cost of drilling.

Well logs represent a specific type of well data obtained by placing sensors in drilled wells. These sensors measure properties such as electromagnetic, radioactive, and acoustic signals, providing continuous measurements along the well trajectory.

In the workflow, these measurements form the basis for petrophysical interpretation, where reservoir properties such as porosity are derived from mathematical and computational methods and prediction algorithms. Following this, static reservoir models are constructed by integrating seismic interpretations with well-based measurements. These models represent the geometry and internal properties of the reservoir using three-dimensional grids. Seismic data provide the structural framework, while well data are used to populate the model with properties such as porosity and permeability. Aligning seismic data in time with well data in depth requires additional processing steps, such as time-to-depth conversion and well

tie-ins, which influence how subsurface structures are represented.

The static model is then used as input to dynamic reservoir models, where fluid flow and pressure development over time are simulated under different production scenarios. These simulations are based on assumptions about reservoir properties and behaviour, and are therefore often evaluated across multiple realisations representing different possible configurations of the reservoir.

The results of these simulations are then used in production forecasting, where expected production profiles and recovery factors are estimated. These forecasts form a central input to economic evaluation, where different development options are assessed. As illustrated in the workflow, this stage ultimately culminates in an investment decision, where uncertainties related to the subsurface representation, forecast quality, and economic assumptions must be considered.

The subsurface workflow is distributed across multiple disciplines, often collectively referred to as "explorationists". These include geophysicists, geologists, petrophysicists, and reservoir engineers. Each discipline contributes a different perspective on the reservoir and works with different types of data and models. Geophysicists primarily work with seismic data to interpret subsurface structures such as horizons and faults. Geologists build structural and stratigraphic interpretations that define the framework of the reservoir. Petrophysicists analyse well logs and core data to estimate reservoir properties. Reservoir engineers develop dynamic models to simulate fluid flow and evaluate production strategies. The outputs from these disciplines are progressively combined as data move through the workflow, linking early-stage interpretations to later-stage forecasts and decisions.

4.3.2 Digitalisation of Subsurface Operations

In recent years, both the scale and pace of subsurface data work have increased. Advances in digital technologies have led to significant growth in the volume of

seismic and well data, reinforcing the data-intensive nature of subsurface activities described above. The challenge is no longer primarily related to data availability, but how available data can be utilised more efficiently and consistently.

This development has been accompanied by increasing pressure to reduce the time required to move from data acquisition to production decisions. Activities such as seismic interpretation, reservoir modelling, and well planning are expected to be carried out more efficiently, while still supporting safe and reliable operations. In this context, efficiency refers not only to speed but also to the ability to work consistently with data across different stages of the workflow.

Within this setting, different approaches are being explored to support more efficient use of subsurface data. This includes the development of industrial data platforms that aim to standardise how data is structured. Such standardisation is also seen as a prerequisite for the effective use of artificial intelligence, where the ability to access consistent and well-structured data becomes central. As outlined in section 4.1, the adoption of such technologies takes place within an environment shaped by safety requirements and regulatory considerations. In practice, these technologies are often positioned as tools to assist existing workflows, rather than replacing them.

As a result, existing ways of working, often based on domain-specific tools and localised data environments, are increasingly challenged by the need to handle larger data volumes and support faster decision-making.

4.4 The Open Subsurface Data Universe (OSDU)

In response to these challenges, the Open Subsurface Data Universe (OSDU) initiative was established as an industry-wide effort to standardise how subsurface data is described and modelled.

OSDU originated from a collaboration between major oil and gas operators,

including Shell, BP, and Equinor, Chevron, Total Energie, and Exxonmobil who identified common challenges related to data management. In particular, data processing and data management were not considered sources of competitive advantage, yet were areas where companies operated largely independently. This led to duplication of effort and limited data sharing. At the same time, subsurface data were often managed within proprietary systems and application-specific formats, which constrained interoperability across tools and organisations.

In response to these challenges, OSDU was established as a joint initiative to develop shared standards for subsurface data. The initiative consists of both an organisational component, the OSDU Forum, and a technical component, the OSDU data platform and associated data models. A central aim of the initiative is to enable a more consistent flow of data from acquisition and interpretation to business decisions.

OSDU Forum

The OSDU Forum, consisting of more than 220 members, serves as the collaborative arena where industry participants, including oil operators, software vendors, and cloud providers, define and agree on standards for subsurface data. Within the forum, participants work to establish common definitions of key data objects, such as wells, seismic surveys, and reservoir models, including their attributes and relationships. These definitions form the basis for a shared data language across the industry. By agreeing on how data should be represented, the forum aims to support interoperability between different software systems and enable more consistent use and exchange of data across organisational boundaries.

OSDU Data Platform

The OSDU data platform represents the technical implementation of these standards. It provides a set of services for storing, indexing, and accessing

subsurface data based on the shared data models defined in the forum. The platform is typically deployed on cloud infrastructure and integrated with existing data environments within oil companies. In practice, the platform acts as an intermediary layer between domain-specific applications and underlying data storage. Data generated within applications can be ingested into the platform and mapped to standardised representations. This enables other applications or users to access the same data through a common interface, independent of the system in which the data was originally created.

OSDU Data Model

The OSDU data models consist of a set of standardised data definitions that specify how different types of subsurface data should be structured. These definitions include both the properties of individual data objects and the relationships between them. In addition to core data content, the models are designed to capture information on how data is produced and used, such as activities, workflows, and data quality attributes.

Chapter 5

Findings

This chapter presents the main findings from our empirical study and data analysis. The findings were developed through the analytical process described in Section 3.3, where we combined a Gioia-inspired data structure with the process-oriented analysis to identify patterns in how subsurface data are contextualised across the workflow. Overall, the findings show that subsurface data are inherently ambiguous and that contextualisation plays a central role in making such data interpretable and usable across contexts.

Our findings are structured into two main sections. Section 5.1 examines the inherent ambiguity of industrial data by showing how data are produced, modelled, and interpreted within and across domain-specific contexts. Section 5.2 then examines how contextualisation is reorganised through the transition toward a shared data platform, focusing on how context is categorised, how data are detached from local environments, and then reconstructed across settings and use-cases, including AI.

5.1 The Inherent Ambiguity of Industrial Data

Industrial data describe phenomena that no measurement fully captures. They are heterogeneous, inconsistently gathered over a long time, incomplete, and thus partial by construction. Their meaning largely depends on the instruments, assumptions, and contexts that produced them. Outside of those contexts, the data alone do not 'speak for themselves', and are therefore inherently ambiguous. The data in the subsurface domain reflect this reality. As a result, producing, representing, and interpreting data about the subsurface becomes about dealing with that ambiguity.

5.1.1 Producing Data about the Subsurface: an Epistemic Problem

Subsurface work is fundamentally shaped by the fact that the reservoir cannot be directly observed. The work involves constructing representations based on limited and indirect measurements. Seismic surveys, well logs, and core samples provide partial and differently structured insights into the subsurface, but none of them offers a complete or unmediated view. As a result, subsurface understanding is built through interpretation rather than direct observation. Only a small portion of the reservoir is ever physically sampled through rock and fluid data from wells. The remaining volume must be inferred from indirect measurements, each associated with its own assumptions and uncertainties. Subsurface data are therefore never simply "given"; they are produced through interpretive processes that shape how the subsurface is represented.

> The different elements have different uncertainties when they have been interpreted, so to speak. We have very few measurements. We have some measurements with seismic and well logs, but the physical evidence we have (...) those are the rocks that we take up. Rock

> samples and fluid samples. That is the only physical evidence we have of what is down there. The rest are measurements or interpretations. But then all of that has an uncertainty that also has to be included in the context of a model.
>
> — Data Manager

Within this setting, early subsurface interpretations often begin with a conceptual geological idea of how the reservoir might be structured. This initial hypothesis — in one petrophysicist's terms, a "geofantasy" — is then progressively evaluated and refined as additional data becomes available. Representations of the subsurface, which later become data themselves, thus emerge through an iterative process in which measurements and geological reasoning are continuously combined.

A key implication of this is that the meaning of subsurface data cannot be separated from the context in which it was produced. Understanding a dataset requires insight into how it was generated, including the assumptions, parameters, and workflows that shaped its interpretation. Contextual information — such as *why* a particular interpretation was made, or *how* a dataset was processed — becomes essential for making sense of the data itself. A data manager emphasised this with an example:

> What is discussed is the context regarding [provenance]. Why was that interpretation made? (...) It is important to know the context in which the dataset was created. Was it to dig a well? Or was it to figure out whether to apply for a license? (...) because when we are planning to dig a well, we have to understand the 'pressure regime' downstream, so that we don't get any surprises. When we dig a neighbouring well, it [can sometimes] happen that we learn that the pressure regime is not the same as the one we referred to in the first well.

In this sense, producing subsurface data is characterised by an ongoing effort to construct meaningful representations under conditions where ambiguity cannot

be eliminated. Rather than attempting to remove this ambiguity, explorationists instead work by acknowledging and explicating it where they can. This makes contextualisation work central to how subsurface data is modeled, and later interpreted.

5.1.2 Modelling Data About the Subsurface: a "Best-Guess" Digital Representation

This epistemic friction is reflected in how subsurface models are constructed and understood in practice. Reservoir models are not treated as objective or final representations of the subsurface, but as provisional accounts that reflect the current state of knowledge. They represent the most plausible interpretation given available data and domain expertise, rather than a definitive description of the reservoir.

> When you create a static reservoir model. It is the best representation or understanding of your reservoir [at all times]. (...) There's nothing that's assumed to be "hard truth", but based on what we know, this is the best we have as of now.
>
> — Petrophysicist

Even when based on the same underlying data, different assumptions about geological structure, connectivity, and properties can lead to multiple plausible models. These alternative representations may imply significantly different outcomes in terms of reservoir volume or production behaviour. As a result, modelling is not oriented toward identifying a single "correct" solution, but toward exploring a space of possible interpretations. To manage this, subsurface work often relies on ensembles of models that capture different, but geologically plausible, scenarios.

> All of this is being modelled, usually in a loop – we call this ensemble modelling – where we modify the input parameters and pick, for each scenario, time interpretation, depth conversion, petrophysical scenarios, facies scenarios, and how you fill these properties (...) So we have the interpretation uncertainty, the time-to-depth conversion uncertainty, and then uncertainties related to how representative the well data are for the entire model space.
>
> — Geologist

The range of acceptable models is actively constrained by, and grounded in, domain-specific knowledge, including geological principles and theoretical understanding. Models must, for example, conform to plausible depositional environments and also flow dynamics. In this way, the solution space is restricted through disciplinary rules that define what constitutes a valid interpretation.

Uncertainty is therefore not reduced by eliminating alternative interpretations, but by structuring and constraining them. What makes a model credible is not that it removes ambiguity, but by explicating it in a controlled and transparent way. This again highlights the role of contextual knowledge, as both the construction and evaluation of models depend on understanding the assumptions and reasoning that underpin them.

These models, however, rely on combining heterogeneous types of subsurface data produced across different disciplines and technical environments.

Reconciling Fragmented Data About the Subsurface

Subsurface models depend on inputs produced across different disciplines and tools, and constructing a model requires bringing these inputs into a single representation. Seismic and well data are the central case. Seismic data represent the subsurface in time and provide continuous spatial coverage; well data are measured in depth and provide high-resolution information at discrete locations.

They do not align directly, and combining them into a coherent representation requires active translation.

This translation relies on velocity models that convert seismic data from time to depth, introducing an additional layer of interpretation. Informants described how multiple scenarios are often used in this conversion, reflecting uncertainty in how subsurface structures are positioned within the model. The datasets also differ in scale and resolution. Seismic data provide broad coverage at relatively low resolution, while well data offer detailed measurements at sparse locations:

> Seismic, for example, at a quite large scale, say, in 2000 meter depth, it's like 510 meter resolution. These are vast volumes of data. You can sample seismic everywhere, but we have an issue of scale, and we have an issue of ambiguity.
>
> Then we have well data that is in a very high resolution, but those are very sparse. So it's just one tiny hole, like a needle in the haystack. So we have high-resolution sparse data and low-resolution dense data covering the entire space. It's a spatial problem. So [we] need to combine those types of data.
>
> — Geologist

Integrating these data types, therefore, involves aligning different representations of the subsurface that differ in measurement domain, scale, and resolution. Ambiguity arises from the assumptions and transformations required to make the data compatible within a single model.

5.1.3 Interpreting Data Within Source System Context

Industrial data are not interpreted in isolation. The systems that store and present them also carry the assumptions, parameters, and prior decisions that make the data intelligible. Working inside such a system, practitioners inherit much of this interpretive context without needing to reconstruct it.

Context Embedded in Proprietary Application Environments

Within proprietary domain-specialised applications, data and representation are tightly coupled to the workflow in which they are produced and the phenomena under investigation. These environments carry their own implicit context, including assumptions, parameter choices, and interpretive decisions. As a result, data is not encountered in isolation, but as a part of an integrated working environment that provides the necessary background for its interpretation.

> They are the data you can trust is correct, and that you can safely use it. The challenge is that, that is only true within the context you are in. Very often, around a specific application. I am working within an application that is specifically made to do production optimisation for a well. (...) that goes to say that, all the data I have within this application is there for a reason. It is there to support production optimisation.
>
> — Data Analyst

This embedding exists because each proprietary application typically maintains a project database that stores the datasets, parameters, and intermediate results associated with an interpretation workflow. These project environments represent the working state of an interpretation and contain rich contextual information about how datasets are produced and used within the modelling process. The application thereby provides the necessary context for interpreting the data within its workflow, even though this context is not inherently represented in the dataset itself.

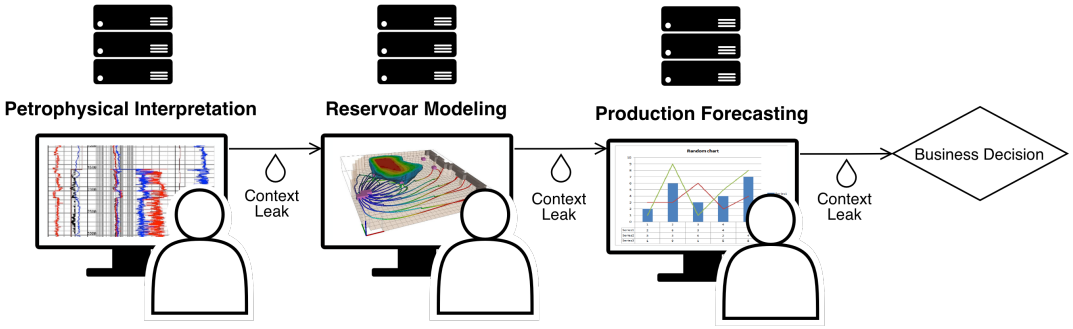

Figure 5.1: Stages of the subsurface workflow (simplified): petrophysical interpretation, reservoir modelling, and production forecasting. Composite figure by the author; layout adapted from Jarvis and Bøklepp (2026).

The embedded context also shapes how practitioners assess whether data can be trusted. Practitioners cannot interpret subsurface data without first being able to trust it, because each interpretation feeds into the next. When data origins and underlying assumptions are unclear, errors can accumulate across workflow stages that only become visible much later, when the cost of correcting them is high. Working inside the application carries a tacit guarantee that the data are fit for the use case the application was built to support, which reduces the perceived need to validate the data further. Trust, in this setting, is generated by the structure of the environment rather than established separately for each dataset.

As a consequence, the trust generated this way is inseparable from the environment that warrants it. It does not attach to the data as such, but to the data as encountered *in situ*. As long as the datasets remain inside the application, both their interpretability and trustworthiness are sustained by the same underlying property: the implicit context of the working environment.

5.1.4 Interpreting Data Across Contexts

With the understanding that industrial data are tightly coupled to their contextual environment, it is worth examining what happens when that data leaves its original context. As data move across contexts, they are interpreted differently, and new

ambiguities are introduced.

Loss of Contextual Information When Data Leaves Proprietary Applications

As data moves across the subsurface workflow, this coupling becomes consequential. Work is organised through sequential handovers between disciplines, where outputs from one stage become inputs to the next.

> It's a bit like a relay race, isn't it? [The] reservoir engineer is dependent on input from a petrophysicist, and we take it from raw data, etc. So you have some loops like that that are repeated, but it's kind of a handover. [And] we have many disciplines.
>
> — Petrophysicist

Each handover does more than transfer data; it also adds new interpretive layers. Disciplines take incoming data and rework it in light of their own assumptions and expertise:

> But the issue, we believe, is that when you have data at every level, [each discipline] take data as input and add their context, their understanding, and their specialised expertise on top, and turn the matured data into information.
>
> — Petrophysicist

These handovers often involve exporting data from proprietary applications into more generic formats such as spreadsheets or exchange files. In this process, the contextual information embedded in the original environment is rarely transferred alongside the data. What remains is a reduced representation, where numerical values are preserved, but the assumptions, parameters, and interpretive reasoning that shaped them are not. A petrophysicist explains:

> We see that the data gets transferred, but what we struggle with is the context... the learning. What were the assumptions? That is not included in the dataset that moves between [applications]. (...) When you are going to build a static model, the tool is, for example, Petrel from SLB. That means the model is stored in Petrel as an SLB-specific type of model. If another application needs data out of it, you have to export it in a typical 3D format, whether that is a CSV file or whatever. And that is often a somewhat dumbed-down version of what sits inside the application database itself.

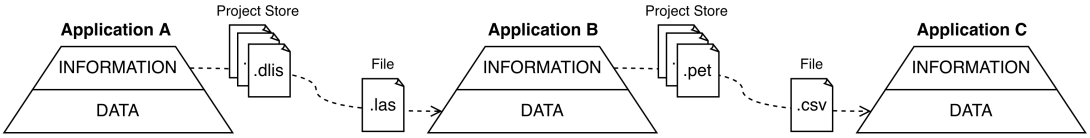

Figure 5.2: Data compression across application boundaries. Rich project-store information is reduced to flat transfer formats (e.g, .csv) when handed off between applications, with progressive loss of context and structure. Adapted from Jarvis and Bøklepp (2026).

As a result, data becomes increasingly detached from the context in which it was produced. Downstream consumers must interpret datasets without access to the underlying assumptions, making it difficult to assess how the data should be understood or how reliable it is. Ambiguity, therefore, arises not only from uncertainty in the data itself but from the loss of contextual information across workflow steps. Figures 5.1 and 5.2 together illustrate the context loss as data moves through the subsurface workflow.

Trust and Misinterpretation Across Contexts

The loss of context also reshapes how trust is established. As described in the previous subsection, data within the context of a proprietary application is trusted structurally: the environment itself warrants the data's fitness for purpose.

However, once data is moved outside these environments, this structural basis no longer applies. Practitioners cannot rely on the application to vouch for the data on their behalf, and in response, they fall back on a relational mechanism. Rather than evaluating the dataset directly, practitioners instead assess who produced it and whether they trust them:

> (...) in real life it is a bit like, when you see a petrophysical interpretation or a geological interpretation, that it is actually such that the people who receive those interpretations... the first thing they check is who did the interpretation? Do they know and trust this person?
>
> — Data Analyst

The recipient's prior relationship with the sender functions by proxy as an assurance of the data's trustworthiness. Trust is thus established through professional relationships and reputation, rather than through the data itself.

For some, this reliance on relational trust is itself a problem. A data scientist described it as a "sickness" in the industry and explained that when he does not trust an upstream interpretation, he simply redoes the work himself:

> (...) I don't want to be negative, but I call it a sickness in our industry. Everyone thinks they're the expert —including me. I always say yes to seismic [interpretations], but [if I don't trust it] we just re-do it. Then I know it is done according to my standard.

Trust, in this setting, becomes relational rather than structural. It no longer rests on the environment in which the data was produced, but on the person who produced it. This shift has consequences for how data is shared and used. For recipients of data, trust depends on familiarity with the producer and their work practices. For producers of data, retaining control over what gets shared,

with whom, and in which version becomes a way of maintaining oversight of the shared understanding of the reservoir model that the collaborators are jointly constructing. This interpersonal way of exchanging datasets is therefore not merely a shortcut for moving data, but also a mechanism for maintaining trust once the structural warrant of the application environment is lost.

Cross-context interpretation problems are not confined to handovers between domain experts. They also surface where data work itself is performed by generalists without domain knowledge. When working with ambiguous data, the challenge is not so much about transforming and moving the data around — what is considered the "data-engineering" part of data work — rather, the challenge is in making sure that data is not misinterpreted or represented inappropriately. Clarifying these misunderstandings requires domain knowledge, yet data work such as ingestion, cleaning, and transformation is performed by generalist data engineers with limited familiarity with the domain. This can create downstream problems that are difficult to detect early and only become visible once their effects emerge.

> If you [mess up] on the technical [aspect], you notice it quite quickly. You know something is not going well: it runs too slow, or the job halts. You sort of get a clear message that "this isn't working", [so] we need to write better pipelines and tests and such. [However], if you are good at [the technical parts] but don't understand the domain, it may look like it works. Everything runs, lots of megabytes are being copied over. But you have misunderstood when transforming from one data model to another what that particular field actually means, what that name meant, and people start using it. And it's only when someone makes an [analytical] report that you start to question [that something looks odd].
>
> — Data Engineer

Establishing trust, therefore, depends on access to contextual knowledge about how the data was produced and what it represents. Without this understanding, errors

may remain undetected until they manifest in downstream analyses or decisions. The data engineer goes on to emphasise that trust is costly to establish but easy to lose:

> If you have spent a lot of time and money to make a data product, and you get [errors relating to context and interpretation]... after two or three times, that trust is gone, and people move on from that data product straight back to the source system, and you have wasted all my effort.

Taken together, these findings show that ambiguity in subsurface data is closely tied to how contextual information is embedded within domain-specific workflows and lost as data moves across disciplines and systems. Contextualisation remains essential for interpreting data, but is largely implicit, localised, and difficult to transfer. As a result, interpretation depends on access to situated knowledge that does not always travel with the data.

In response to these challenges, there is an ongoing shift toward making contextual information more explicit and accessible across the subsurface workflow. Rather than relying on context embedded within individual applications and work practices, emerging approaches seek to represent and manage context through a shared data platform. This marks a transition in how ambiguity is addressed, where contextual information is increasingly externalised and structured to support interpretation across domains.

5.2 Contextualising Industrial Data in a Shared Data Platform

Where section 5.1 examined how industrial data become ambiguous through the conditions of their production and the contexts in which they are interpreted,

this section turns to how that ambiguity is addressed when data move onto a shared platform. Contextualisation in this setting is no longer carried implicitly by source systems and disciplinary practice. The platform must define data structures explicitly, support the work of moving data into them, and reconstruct the contextual grounding that source systems previously provided. The Open Subsurface Data Universe is one such platform that attempts this, and its development at Equinor offers a setting in which to examine how this work of explicit contextualisation is undertaken in practice.

5.2.1 Categorising Data in Shared Data Models

Before data from different sources can be worked with together, the platform has to fix the categories and relationships in which all of it will be expressed. This definitional work is encoded in two layers: shared data models, which specify the kinds of entities that exist on the platform, and metadata structures, which describe how individual datasets relate to those entities and to one another.

Standardisation of Data Models

Interoperability depends on agreement about fundamentals. What counts as a well? What attributes and relationships must it carry? What makes a measurement a well-log rather than something else? Shared data models answer these questions in a form that holds across disciplines and formats. Without this prior agreement, the same dataset can mean different things in different parts of the organisation, and cross-context interpretation collapses back into the disciplinary and application-specific silos that the platform is meant to bridge.

Standardisation at this level is also a shift in who defines the categories. What counts as a well or a well-log on the platform is no longer determined by individual applications or legacy formats, but negotiated collectively by the consortium that maintains the platform. Definitional authority moves out of the source systems

and into a shared layer.

This negotiation has consequences for what the data models look like in practice. Operators bring different requirements based on the applications they run, and accommodating all of them within a single shared schema would either force every dataset to carry every possible attribute or restrict the platform to a narrow common denominator. The compromise is to make most attributes optional:

> In OSDU, 99% of all data objects are optional. (...) [Because] the challenge is that one group says "I need A, B, and C, and I want them as requirements on OSDU." And then I talk to the next group, who also know the domain and have other requirements, and a third group with their own requirements, and before we know it, you need everything. Suddenly all kinds of small details have become mandatory because someone has said "I need this" in their context.
>
> — Data Analyst

The result is a data model architecture organised around a minimum viable representation: an object can be registered on the platform with very little information, and additional attributes are populated as use cases demand them. Because the platform tolerates sparse objects, the work of making data usable falls on the applications that consume it and on the people preparing data for specific workflows.

Structural and Provenance Metadata

Beyond defining categories, the platform attaches metadata to each dataset. Structural metadata makes the relationships between datasets explicit: identifiers, spatial references, and links to other entities allow a dataset to be located, connected, and reused across systems. A well-log connects to a well, a well to a field, a field to a seismic survey. These relationships were previously held together

within specific applications or by practitioner knowledge; the platform represents them directly.

A second form of metadata captures how and why a dataset was produced — the workflows, parameters, and assumptions that shaped it. These were previously implicit in the modelling environment. The platform makes them part of the data structure:

> The activity [model] can describe what business activity lies behind that dataset here. (...) It can also describe in detail how the dataset was created, [such as] what parameters were used in the algorithms to create the dataset.
>
> — Data Manager

Together, structural and provenance metadata make data partially interpretable outside the environment in which it was originally created. What was previously embedded in tools, workflows, and expert judgment is now encoded in the data model. The work involved in categorising the data as such sets the conditions for sharing and moving the data out of its original setting.

5.2.2 Detaching Data From Their Original Environment

Detachment is the separation of data from the context they were produced in. In the case of OSDU, this means moving data out of the environment around its source systems, where what those systems previously warranted — interpretive context, fitness for purpose, and lineage — does not follow. OSDU attempts to provide explicit substitutes for each: Activity Templates and Activity Models for the purpose data was produced, Technical Assurance for how fit it is for that purpose, and lineage for its history of modification. These substitutes detach the data from its original context, making it available for the work of reconstructing context that follows.

Activity Templates and Activity Models When a dataset leaves its source system, the workflow that produced it does not travel with it. The platform provides an explicit substitute through two related objects. An Activity Template defines a business process generically, while an Activity Model records what was done in a specific instance, including the data used, the parameters chosen, and the people involved:

> [OSDU has] an object called 'Activity Template'. (...) [Where you] define an activity — a business process. For example, the whole process from acquisition of raw logs to delivery of an interpretation. From raw logs to quality evaluation is one step. From quality evaluation to a composite is another step. And then you can sew it all together in a petrophysical workflow. Then you reference exactly the data elements that are used. And [in the Activity Model] there are timestamps and person... "Completed by petrophysicist "Espen Askeladd" on 12.12.11".
>
> — Data Analyst

Smaller activities, such as quality evaluation and composite construction, can be assembled into larger workflows like petrophysical interpretation. Each step references the specific data elements it consumes and produces, along with the timestamps and people involved. What previously lived implicitly inside an application — what the data was for, how it was generated, by whom, and when — becomes explicit and attached to the data itself.

Technical Assurance Activity Models attach data to the workflow it was produced for, but they do not say how well the data serves that workflow. Where this assessment was previously warranted through interpersonal exchange or localised procedure — practitioners trusting senders, applications guaranteeing fitness for their built-in use cases — the platform has to make it explicit.

Technical Assurance is the mechanism that does this. Drawing on OSDU's documentation, it "describes the record's overall suitability for general business

consumption based on data quality" and "... provides a standard set of reference values that describe various levels of quality and confidence in the data" (OSDU, 2024). The values range from certified to unsuitable, attached to the dataset as metadata that any downstream user can read.

Establishing a Technical Assurance value requires evaluating the data anew. Source systems used to carry fitness implicitly. A dataset inside a production-optimisation application was, by being there, considered fit for production optimisation. The platform asks practitioners to make that judgment differently, in terms of a generic workflow rather than the application that produced the data. Fitness for purpose is detached from the system that previously vouched for it, and re-expressed as an explicit indicator on the data itself.

Lineage and Versioning The third aspect that does not survive detachment is the dataset's history. Inside a source system, a practitioner working with data could ask the person who sent it which version they were looking at, what assumptions had gone into that version, and whether anything had been changed since. On the platform, where data is no longer exchanged interpersonally, those questions have to be answerable from the data itself.

OSDU addresses this by versioning every record. When a record is created, it is automatically assigned a version, and every subsequent modification produces a new immutable version that exists alongside the previous ones. A record that has been modified four times exists on the platform as four versions, each individually addressable.

For a petrophysicist this makes new kinds of work possible:

> Then you have the possibility to run many different petrophysical interpretations because you have control over the versions and how data is handled. So you can say that you have an interpretation as it is today, that for example [is optimised] to focus on a [specific] well. But then it might be that you want to run a different kind of interpretation,

> like a field standard for example, [so that] all the wells on that field get interpreted in the same manner.

Versioning is currently implemented within OSDU but not across systems. Informants describe this as a known limitation and speak optimistically about cross-system lineage as a future capability — one that would let the history of a dataset be traced back through the systems and transformations that produced it. Data exchanged outside the platform, through email, SharePoint, or personal contact, can be effectively lost over time:

> I don't know how many you have talked to who have said that they get their data via e-mail and Teams and SharePoint and so on. We have to get rid of that. [When they ask me for data] I have no idea, because I got it on e-mail ten years ago. But those are actually the data we are now trying to find again. I get questions like that. I worked on a project, and someone says, "I see an e-mail here. Have you seen that data?" [laughs] "No, I don't think so..."
>
> — Data Scientist

What versioning does for detachment is shift the basis on which a dataset's identity is established. Inside a source system, data was implicitly the current state of the work, and questions about prior states required asking the person who held it. The platform makes the dataset's history part of its representation instead. Questions that previously depended on knowing the sender ("which version is this? what changed?") become questions the dataset itself can answer.

5.2.3 Reconstructing Context Across Domains

Detachment opens up the possibility of using data across domains and purposes beyond the one it was originally produced for. However, that first requires the data to be re-contextualised: tied to new workflows it can serve, evaluated for

those workflows, and traced back to the systems and decisions that shaped it. The mechanisms set up in the previous section make this work possible, but they do not perform it. Reconstruction is what users and data workers do with those mechanisms once data is on the platform.

An Activity Template does not describe a specific workflow until it has been populated as an Activity Model. The template defines placeholders for inputs, parameters, methodology, outputs, and the people involved, but these remain empty until a practitioner connects them to the concrete datasets and documents what was done. Until then, the model carries no context. A data analyst put it directly:

> The network of connections doesn't come for free. OSDU is not something where you can take some brown lumps and throw them into the grinder called OSDU, and then golden nuggets come out the other side. Someone actually has to contextualise. Someone has to connect all these data points together.
>
> — Data Analyst

The labour of re-contextualising data is experienced as an additional burden, but it is also what makes the data trustworthy on the platform. Establishing a Technical Assurance value signals that someone has examined the data and judged it against the platform's requirements. A data scientist described this as more of a feature rather than a drawback:

> So that's a good thing about OSDU. We know we have control on [the data]. We have people who have uploaded it, looked at it, and if there perhaps were any errors in it [we know] that they have used their expertise. So we can always trust that the data we use is safe. Of course there could be errors, but we have processes to look after that.

What was previously the implicit warrant of an application is reconstructed through the visible work of practitioners who pass the data through the platform's mechanisms.

In regard to lineage, being able to trace a dataset back to its origin provides a way of understanding how it should be interpreted. One informant emphasised this using the metric of pressure as an example:

> (...) because when we are planning to dig a well, we have to understand the 'pressure regime' downstream, so that we don't get any surprises. When we dig a neighbouring well, it [can sometimes] happen that we learn that the pressure regime is not the same as the one we referred to in the first well.

Through this, as well as our previous example regarding the multiple ways of handling the same well data from the earlier subsection, highlights how provenance does not only document the history of the data but also supports its interpretation in new contexts. By linking data to the conditions under which it was produced, users gain a basis for assessing what the data represents and how it can be used.

However, the shift to a shared data platform also introduces tensions in this regard. While the platform enables data to be shared beyond established collaborations, it also reduces reliance on informal exchange between colleagues. Practices such as sharing data directly via email may bypass the lineage and versioning mechanisms established in the platform, thereby disrupting the shared frame of reference that these structures provide. Hence, the transition to a shared data platform is not a smooth replacement of existing practices. Rather, new and old ways of working coexist, and may at times pull data out of the structures that are intended to support its use across contexts.

In this sense, contextualisation is not completed when data is represented in the platform, but must be carried out again in use. While the platform makes contextual information available through provenance and traceability, users

still need to relate datasets to their specific task and domain. Rather than eliminating ambiguity, the shared data platform redistributes it, shifting the work of contextualisation from being embedded in local tools and practices to becoming a more explicit and ongoing process across domains. At the same time, this work reshapes how context is represented in the platform. By linking datasets to their origin, production processes, and relationships to other data, contextual information becomes part of how data is described and organised. This makes it possible to access and use data beyond the purposes for which it was intended.

5.2.4 Interpreting Contextualised Data for AI

As shown in the preceding sections, the shared data platform makes subsurface data available beyond the contexts in which it was originally produced. Informants emphasised that this also opens up for new forms of use, including the application of AI to large collections of data. In this setting, AI provides a way of examining what is required for data to be used beyond its original context. The same mechanisms that enable data to be interpreted across domains - by making contextual relationships explicit and traceable - also make data available for AI-based analysis. In this way, the shared data platform extends the work of contextualisation by requiring it to be represented in standardised and machine-readable forms.

Informants emphasised that while large amounts of subsurface data already exist to be used in AI applications, this data is not readily usable for such purposes. In earlier work, data from sources such as Diskos had to be manually gathered and curated before it could be used. In contrast, OSDU provides a way of consolidating data and quality assuring it as part of the platform, making it possible to use it more directly for further AI applications:

> So, why do we need OSDU? We have all [the] data in Diskos. If you want to use AI (...) Then you need large amounts of data that (...)

> which in any case must be somewhat quality assured. Here you have data that you insert, quality assure it. And then you can use it for an application [for example] to create a foundation model.
>
> — Data Scientist

This highlights that making data available is not the only factor for efficient use of AI. Data must also be prepared in ways that assure its quality and consistency across datasets. The shared data platform is therefore not only a repository, but a setting in which data is actively structured and qualified for further use.

In addition, it was emphasised that when connections between wells, fields, trajectories, and historical data are made explicit, data can be processed as part of a broader structure rather than as isolated measurements. This enables AI systems to operate on relationships, not only individual values.

> If you contextualise the data, there is an incredible amount more for an AI to work with (...) a well path is connected to a well, which belongs to a field, which has a rig, trajectories, and historical data. You get such an extreme network of things that an AI can start to delve into.
>
> — Data Analyst

However, making such relationships explicit does not ensure that they are complete or sufficient. Contextual information that remains implicit, uncertain, or unevenly represented may not be captured in the data that AI systems rely on. As a result, AI may process data without accounting for how measurements depend on specific geological or operational conditions.

> (...) You have had startups that have looked at all the data released by NPD, Diskos, and the database system. There, we have done some AI to find Bypass Oil, that is, the area where we have overlooked oil. And there I got a presentation where they boasted that yes, "there I have

> found, and there I have found." But there I was as a petrophysicist, I could only say no, but it is wrong because [in that field] we have the [data] quality like that. It is completely obvious what has happened. So it is this context or the one combined with [the data] quality, and what the other data means. Even if you have a good measurement, you are not measuring what you think you are measuring.
>
> — Petrophysicist

From their perspective, it was “completely obvious what had happened”: the system failed to account for how data quality and contextual information shape what a given measurement actually indicates. Rather than resolving ambiguity, AI can therefore reproduce or amplify existing uncertainties when contextual relationships are not adequately represented.

These limitations are not only technical, but also reflect more fundamental properties of subsurface data. As each reservoir is shaped by distinct characteristics, data cannot be assumed to be directly comparable across contexts.

> So every well drill hole is different, every reservoir anyway, completely different. Everything has its own characteristics. I mean, different depth, deposited from different areas, different rock types, different time spaces. When it gets deposited, it means you have a subsurface that subsides or moves in time. So that’s why it is really difficult to use AI.
>
> — Geologist

Even when data is structured and contextualised, the use of AI remains constrained by how decisions are made in practice. In safety-critical settings, such as subsurface work, outputs are not treated as decisions, but as suggestions that domain experts must evaluate. This reflects the consequences of incorrect interpretations, where errors may have significant operational and financial implications.

> We are a bit too conservative. It has to do with the industry historically, but it also has to do with the consequences of what can happen if things go wrong, right? (...) It's quite far-fetched to let a machine just suggest a well (...) There's a high degree of conservatism in that. And there may be a lot of things that we use AI or ML to suggest, but none of that works. It just becomes a kind of suggestion.
>
> — Petrophysicist

Following the example of well-planning, decisions need to be based on reasoning that can be traced and justified. This limits the role of stochastic models that produce varying results from the same input, as such variability introduces ambiguity at the level of the output. This leads to a cautious approach in which AI supports, rather than replaces, domain experts in their work of interpreting subsurface data. In practice, the use of AI requires it to be aligned with existing work processes.

> One thing we're really concerned about is whether we can safely use it in our processes. We can't just ignore processes that we already have. And what we've discovered [is that] Generative AI models are pure guesswork. That's not good for those of us working in the subsurface. Where you want to bore a well (...) It has to be planned safely. So we have to have the same answer every time. It starts with having data that is clean, that we have control over.
>
> — Data Scientist

Making data usable for AI, therefore, involves preserving the conditions under which the data is meaningful and safe to use under existing work processes. This requires ongoing work to define how data should be structured, what relationships should be made explicit, and how differences in interpretation and quality should be accounted for. Consequently, AI does not operate on self-evident data, but on data that has been actively prepared for analysis. The effectiveness of AI is thus

closely tied to how this preparatory work is carried out. Rather than reducing the need for interpretation, the introduction of AI shifts and expands it, making the work of structuring and relating data, as well as assessing its quality, more central across the subsurface workflow.

Chapter 6

Discussion

This thesis aims to answer the research question: *How is industrial data disambiguated through contextualisation as it departs from its origin and is reused in other settings?* Our central contribution is a process model of how data are disambiguated through data contextualisation. The model supports two claims. First, disambiguation does not proceed in a single direction: data's ambiguity is not resolved once and for all but becomes more and less pronounced as data are detached from their origins and reconstructed for new uses — making ambiguity something practitioners continually manage rather than eliminate. Second, the interpretive work that makes data usable for analysis, including AI-based analysis, shifts upstream. Rather than being concentrated downstream in making AI outputs meaningful after prediction, contextualisation increasingly takes the form of continuously reconstructing the contextual conditions of the data themselves before analysis — work that stabilises data only provisionally for a given purpose rather than resolving its ambiguity for good.

This section is structured as follows. First, we present our process model and account for where and when it applies. Then, we discuss the model as it relates to relevant literature and offer our contributions, as well as highlighting its practical implications.

6.1 A Process Model for Disambiguating Data Through Context

The process model comprises three subprocesses — *categorizing*, *detaching*, and *reconstructing* — that carry data through four states: *embedded*, *codified*, *generic*, and *recontextualised*. Each subprocess moves data between adjacent states and shifts the level of ambiguity in a different direction. The arrangement is not a linear sequence with sharp boundaries. The subprocesses blend into each other, and the process is recurrent: reconstructed data feeds back into downstream uses, including the source systems in which the cycle began. Across the three subprocesses, ambiguity is first reduced, then increased, then reduced again, but in different registers and through different work each time.

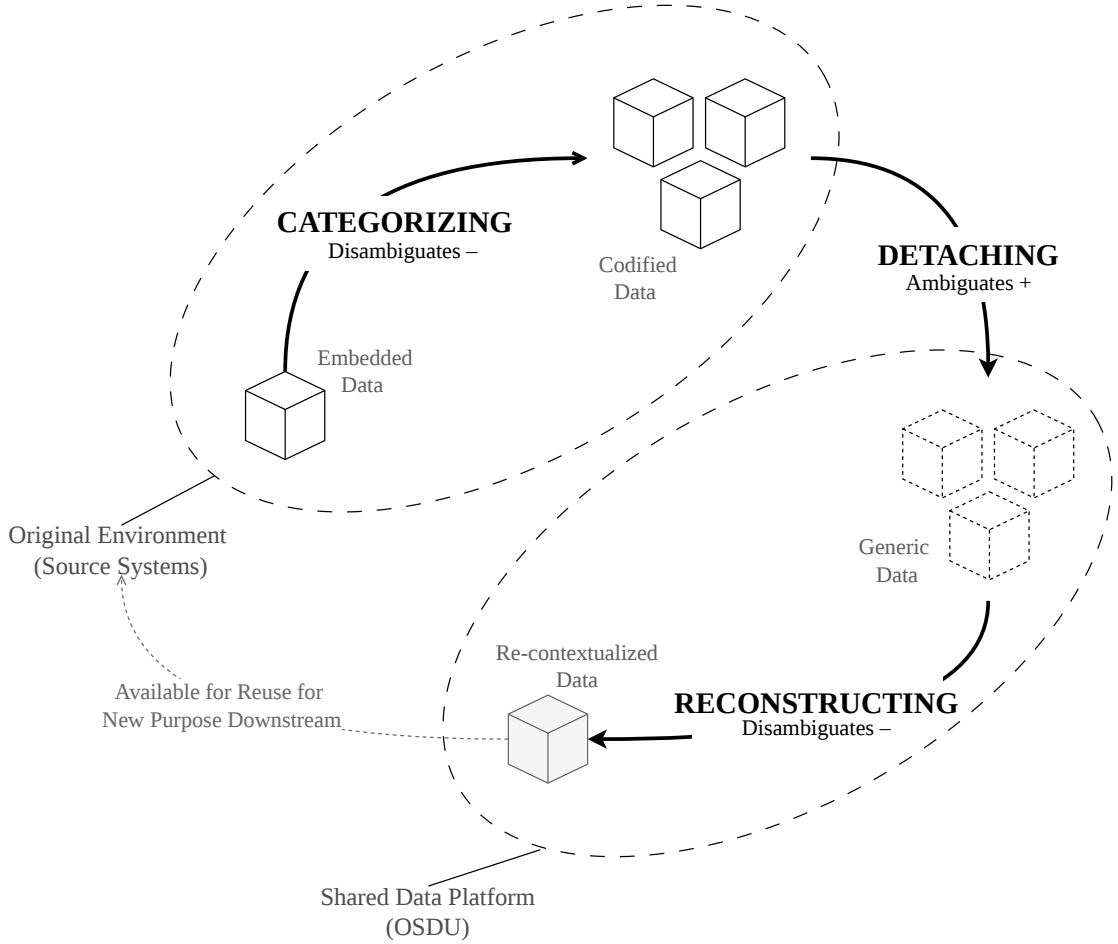

Figure 6.1: A process model for disambiguation through contextualisation. Starts at 'Embedded Data' and follows through the core processes of contextualisation. Dashed return arc indicates that the process is ongoing and recurrent, rather than linear.

The starting point is data *embedded* in the source systems that produced them. In this state, data are inseparable from the conventions of the discipline that generated them, the assumptions of the workflow that produced them, and the application environments that hold them. Their interpretability is local and largely tacit: a practitioner working inside the same environment recognises what the data presuppose, but that recognition is not carried by the data themselves. For instance, working within an application made for production optimisation, there is a tacit understanding that the data used as inputs are appropriate for production optimisation. The data does not necessarily need further context, because the environment warrants its fit for purpose. As such, data in the *embedded* state may lack context to be used across other settings, making it inseparable from its

environment.

Categorizing. The first subprocess externalises the categorical and relational apparatus that previously lived inside local applications and expert practice. For instance, in our case, data models that are defined collectively in OSDU determine what counts as a "well", what attributes a well log carries, and how these objects relate to other objects in the domain. This goes beyond just labelling. The data models specify what entities exist in the domain and how they relate, so data are placed within a structure rather than just tagged. Once data are attributed to these models — well logs treated as well data, surveys tied to wells, and so on — they enter the *codified* state. They may still sit in the source systems, but they are now also addressable through shared categories and APIs that make their structure explicit beyond the application holding them. This is a partial form of disambiguation. Categorising defines what an object is and how it relates to other objects. It does not capture the assumptions, parameters, and methodological choices that produced the values. It defines, for example, that this well log is tied to that well, and that it broadly belongs to the category of well data, among others.

Detaching. To make data useful across contexts, data must eventually move out of the source systems whose conventions and applications shaped them. When data move out of those systems, they lose the embedded scaffolding that made them interpretable. The bytes that make up the numerical values of a well log can easily be transferred from its source application onto a platform. However, assumptions about what the log was optimised for, the acquisition conditions, and the corrections done in the source environment cannot. Data that is moved out of its source systems, therefore, enters the *generic* state *more ambiguous* than before: structurally placed by the categorisation step, but stripped of the interpretive context that anchored them to any particular use. For example, a well log moved out of its source environment still carries the enrichment from the previous categorisation step; it is still a well log that is tied to so-and-so well. What is missing is the context of the environment for which this use case was appropriate.

Detaching is also when preparation for later reconstruction begins. In our empirical case, we identified these as three procedures on the OSDU platform: technical assurance, activity templates, and lineage systems. Practitioners will draw on these later, in the reconstructing subprocess, when generic data are reconnected to a new purpose.

Reconstructing. Reconstruction is the work of attaching data to a new interpretive context. In our case, practitioners do this by tying the data to a specific workflow and the inputs to the assumptions that frame what is being modelled. This involves concrete connections — linking a well to a field, fitting a log to a particular reservoir model. The procedures set up during detaching come into use here. Activity templates provide a workflow structure. When practitioners populate an activity template with detached data, it becomes a concrete activity model. Technical assurance lets practitioners re-evaluate data quality against the demands of the new purpose, which is often different from the purpose for which the data were originally produced. Where lineage systems are operational, they surface the history and transformations the data carry, so practitioners can examine provenance directly. Where they are not, this work falls to people: someone has to reconnect the data to its history, the assumptions that went into producing it, and the conditions of its production. Reconstructing also disambiguates to a degree, but the work is different from categorisation. Categorisation places data within a shared structure. Reconstructing fits that data to a specific activity and the evaluation that the activity requires.

Once data has been reconstructed for a new purpose, it becomes available for reuse downstream. In many cases, this includes feeding back into the source systems and workflows from which the cycle began: a petrophysical interpretation reconstructed for production optimisation may inform well planning; data reconstructed for a reservoir model may feed the next round of seismic interpretation. The dashed return arc in Figure 6.1 marks this loop. The process is recurrent rather than linear: the same data, or data derived from it, may move repeatedly through categorising, detaching, and reconstructing as it is produced,

qualified, and put to new uses.

Scope Conditions

The process model is intended for industrial data that carries the residue of substantial prior interpretive work and lacks a directly accessible referent. Subsurface data is a clear instance: reservoirs cannot be observed directly, and every dataset is several inferential steps removed from the phenomenon it describes. The model also presupposes that the apparatus the sub-processes depend on is in place — shared data models for categorisation, and substitutes such as Activity Templates, Technical Assurance, and lineage to carry data through detachment and reconstruction. Where these are absent, the cycle does not disappear but runs entirely through people. Where data instead has a stable and directly observable referent, much of the disambiguation work the model describes becomes redundant.

6.2 Theoretical Contributions

6.2.1 Disambiguation as a Non-Linear Process

Our process model shows that the disambiguation of subsurface data is not linear. Disambiguation happens, is undone, and is then done a second time — through different work each time. Categorising disambiguates by externalising embedded data into shared and commonly understood data models. Detaching increases ambiguity, because the moment data are separated from the source systems whose conventions and applications anchored their meaning, the assumptions that made them interpretable are no longer carried with them. Reconstructing then disambiguates again, but in a different register, by re-fitting detached data to a new activity model, a new purpose, and a new evaluative apparatus such as technical assurance. On this account, disambiguation is achieved twice and through different work each time, with a deliberate detour through ambiguity in between — though

each round is partial, stabilising data for a purpose rather than resolving the inherent ambiguity of data whose referent cannot be directly observed.

This trajectory connects to a wider conversation in IS about how data is acquired and retains meaning. A common claim across this literature is that representational faithfulness is not a property data possess but a quality they acquire through ongoing work, and that the conditions sustaining that work can shift. Østerlie and Monteiro (2020) makes this point through sensor data in offshore production: representations that hold under one set of operating conditions can stop holding under another, so the data work that maintains them is part of what the representations are. Mikalsen and Monteiro (2021, p. 25) extends the picture to interpretive work under uncertainty. In their study of offshore explorationists, "the crafting of institutional facts is not a steady march from uncertain, error-prone data to solid facts," but a process in which "epistemological uncertainty is provisionally bracketed to reach a decision" while never being eliminated. Their three patterns — accumulating, reframing, and prospecting — describe a fundamentally abductive practice in which interpretations are forged from under-determined data, contested when incompatible new data arrive, and held in tension with competing alternatives. Our findings build on these insights: across subsurface workflows, practitioners do the kind of continuous, abductive interpretive work these accounts describe.

However, our model identifies an additional structural feature of this work — the detach-to-reconstruct seam. As illustrated through our process model, prior interpretations cease to constrain the data, and practitioners must abductively search for an interpretation that fits a new context. This is structurally analogous to Mikalsen and Monteiro (2021)'s reframing pattern, but with a different trigger. In their case, reframing is provoked by the contingent arrival of incompatible new data. In ours, it is built into the routine movement of platform-mediated data, which detaches data from the contexts that disambiguated them and demands their reconstruction as a matter of course. The work of producing institutional facts under uncertainty, on our account, is not only continuous but oscillating:

data cycles between phases where it is more and less ambiguous depending on where it sits in the platform trajectory, and the oscillation is structural rather than contingent.

Moreover, our study also contributes with insights regarding how the cycle of detaching and reconstructing neither recovers the original context nor is incidental to portability, but is itself the work that makes data portable. Where Aaltonen et al. (2021) describes a cycle in which data lose context as they are made portable and regain it when recontextualised, our model adds two specifications. First, the regained context is not the same as the one lost. It has been fitted to a new activity, evaluated against a different standard, and traceable through different procedures. Reconstruction, in this sense, is not the recovery of an objective referent but the partial reassembly of an interpretive context that may itself have been transformed in transit. Second, where that account treats recontextualisation as a recovery step appended to portability, our model treats the categorise–detach–reconstruct cycle as the mechanism that produces portability in the first place. Each subprocess in the model does distinct work, and the explicit substitutes set up during detaching — Activity Templates, Technical Assurance, lineage — resemble what Parmiggiani et al. (2024) call anticipatory curation. They are placed in advance against an interpretive context that has not yet been encountered.

Overall, our process model contributes to the literature on data in IS by emphasising the non-linearity of the process of disambiguation. We show how data becomes more ambiguous as they are detached from their origins, but is further disambiguated when contextualised for new purposes. In our process model, this is evident through the processes of detaching and reconstructing. Existing accounts have shown that data work is ongoing, abductive, and conditional on context that cannot fully be made portable (Aaltonen et al., 2021; Mikalsen & Monteiro, 2021; Østerlie & Monteiro, 2020). Our central contribution is to specify what the work of reconstruction actually does: it does not recover or replicate the data's original context, but deliberately makes the data reusable in a new one. Reconstruction is therefore not a maintenance operation that keeps representations faithful to an

original referent, but the productive work of re-anchoring data to a context it was not produced for.

6.2.2 Recontextualising Data for AI

Our process model shows that contextualization disambiguates data, preparing it for downstream interpretation, including those made or supported by AI systems. Through the transition to a shared data platform, the findings identify that much of this preparation work increasingly shifts upstream. Rather than interpreting and adding meaning to AI outputs– essentially ccontextualising them, this work instead happens through the work of structuring, relating, and qualifying the data before they are used in other such settings. Thus, contextualisation becomes structurally part of the process required to make data usable for future analytical use, including AI-based analysis.

Our findings build on prior studies showing that AI systems do not produce knowledge independently of human expertise, but rely on ongoing work around data, interpretation, and validation (Lebovitz et al., 2021; van den Broek, 2025; van den Broek et al., 2021; Waardenburg & Sergeeva, 2021). Rather than operating on data that is considered "raw", as the AI paradigm presupposes (van den Broek et al., 2021), AI systems depend on data whose meaning and relevance must continuously be established in practice. Consequently, producing meaningful AI outputs becomes multifaceted and iterative, rather than a linear process from data to knowledge (van den Broek, 2025). In the case of OSDU, this becomes visible through the work required to structure, qualify, and relate subsurface datasets before they become reusable across downstream analytical settings.

However, while prior studies primarily examine how experts make AI outputs meaningful after prediction has occurred (van den Broek et al., 2021; Waardenburg & Sergeeva, 2021), our findings show that contextualisation increasingly occurs earlier in the process, at the level of reconstructing the contextual conditions of the underlying data themselves. Waardenburg and Sergeeva (2021), for example,

highlights how algorithmic brokers contextualise opaque predictions to make them meaningful and trustworthy for end-users, while van den Broek et al. (2021) emphasises the role of situated expertise in interpreting AI-based hiring recommendations. Lebovitz et al. (2021) extends this account by shifting the focus towards the "ground truth" labels of such predictions. In domains characterised by uncertainty, AI systems need to be aligned with experts' "know-how", due to the underlying ambiguity of experts' situated practices where knowledge claims are subject to multiple interpretations (Lebovitz et al., 2021). Our findings resonate with and extend this literature by showing how contextualisation increasingly shifts upstream in platform-mediated workflows, where interpretability depends on continuously reconstructing the contextual conditions that allow data to remain meaningful as they move across domains and analytical settings.

This shift is evident in the repetitive work required to establish relationships, to obtain high-quality datasets, and to maintain provenance across workflows. Interpretability no longer relies primarily on tacit familiarity with local workflows and proprietary application environments, but increasingly depends on explicit and traceable contextual representations embedded in the shared data platform. In this way, the transition to OSDU illustrates van den Broek (2025)'s argument that AI initiatives intensify and redistribute data work by reorganising professional practices around the ongoing production and maintenance of machine-readable data. As contextualisation becomes embedded through platform features such as shared data models, lineage tracking, and technical assurance, sustaining interpretability increasingly depends on the continuous work by domain experts of making contextual conditions explicit and traceable.

This also reflects the increasingly distributed nature of AI-related data work, where the preparation and contextualisation of data become organisationally separated across actors, settings, and points in time from their eventual use (van den Broek, 2025). As subsurface data are detached from their original settings and reused for AI-based analysis, responsibility for preserving their meaning becomes redistributed across the workflow. Rather than remaining confined to

data analysts or curators, contextualisation becomes embedded into the ongoing work of domain experts, who must continuously relate, qualify, and reconstruct data for new analytical settings. Our process model further develops van den Broek (2025)'s argument by showing how the redistribution of AI-related data work also reorganises where contextualisation occurs in practice, increasingly shifting interpretive work upstream into the preparation of the data itself.

Furthermore, our findings show that AI systems reproduce or amplify underlying ambiguity when data are reused outside the contexts in which their meaning was originally established. This becomes especially challenging in subsurface work, where measurements cannot be assumed to retain the same meaning across geological and operational settings. This resonates with Felin and Holweg (2024)'s argument that data cannot independently identify what counts as relevant evidence and continue to rely on theoretical interpretation in uncertain domains. In our case, practitioners described how AI-based analysis operated on measurements detached from the geological and operational conditions that originally gave them meaning. Although the measurements themselves appeared valid, practitioners argued that the system failed to account for the contextual relationships necessary to interpret what the data actually indicated. Consequently, the findings support Felin and Holweg (2024)'s argument that AI systems lack the forward-looking causal reasoning needed to determine whether data are actually relevant to a novel situation.

In light of this, we build on Lebovitz et al. (2021)'s argument that "ground truth" cannot function as a stable mirror of reality. Data does not faithfully represent reality, but operates as signs that stand in for phenomena and therefore require ongoing interpretation (Alaimo et al., 2020; Mikalsen & Monteiro, 2021). Rather than resolving ambiguity once and for all, contextualisation temporarily stabilises data for particular downstream purposes. In addition, we reinforce van den Broek (2025)'s point that AI systems depend on ongoing work to establish the relevance of data in practice. Our findings show that AI does not reduce the need for interpretation, but intensifies the reconstruction work required to establish whether

data remain meaningful and relevant across contexts.

Overall, our process model contributes to the AI-at-work literature by showing how the transition to a shared data platform reorganises contextualization work. Instead of being centred on evaluating algorithmic outputs, this work shifts upstream, making data available for reuse and AI-based analysis to begin with. More broadly, the study shows how data are disambiguated as they move across settings, where data increasingly depend on the ongoing work of categorising, detaching, and reconstructing contextual conditions that make them meaningful and reusable in new settings. The context of AI use becomes a particularly revealing case of the broader challenge of sustaining interpretability as data depart from their origin and are reused across domains and analytical purposes.

6.3 Practical Implications

What follows are suggestions for what organisations should consider in their pursuit of disambiguating industrial data within or across contexts.

First, the interpretive work that disambiguates industrial data is largely done by domain experts in the course of their normal practice. This includes, for example, noting what assumptions a model rests on, flagging when data looks anomalous, passing on what was decided and why. Much of this work is invisible until something downstream depends on it. Expecting it to happen in the gaps between deliverables leaves the data without the context that made it usable in the first place. Organisations should build contextualisation into roles, time, and tooling. Write it into expectations and give it space in workflows. Ideally, capturing the interpretive moves at the point they are made, instead of weeks later from memory.

Second, a platform like OSDU can carry structure, identifiers, lineage, and the kinds of provenance that fit into a schema. It cannot, however, carry the tacit interpretive judgement that domain experts apply when they read a log or weigh a model. Organisations that treat the platform as a substitute for application-

embedded context will produce datasets that look ready to use but are missing the conditions under which they were trustworthy. The platform should be planned alongside the channels through which those judgements are made, and old practices should be expected to coexist in parallel with the new ones throughout the transition. And because contextualisation does not move data on a straight line from ambiguous to disambiguous, the work of fitting data into shared models, evaluating it, and recording who did what is not an overhead cost to be minimised. It is where much of the disambiguation actually happens.

Finally, centralising data on a shared platform is not enough for effective AI use. AI systems work on data whose history, relationships, and quality can be read consistently across contexts, and that reading depends on context travelling with the data. In practice, much of it does not. Datasets get shared through Teams, email, and SharePoint, and the context falls off along the way. A log file arrives without the assumptions that went into producing it. A model result built from it circulates without the conditions under which it was valid. Organisations pursuing AI on industrial data should treat that contextual information — who produced the data, how, and under what conditions — as part of the dataset itself, not as commentary that lives in inboxes and chat threads. AI readiness is not primarily a question of data volume, but of whether the meaning of that data can be reconstructed when it is needed.

Chapter 7

Conclusion

This thesis has aimed to shed light on how industrial data is disambiguated through data contextualisation, particularly regarding when data departs from its origins and is reused in other settings. It draws on a body of research on data, data work, and adjacent AI literature. The study is a qualitative and interpretive single-case study in design, and has studied an oil and gas operator on the Norwegian Continental Shelf through its ongoing transition to a shared data platform. The study demonstrates in which ways contextualisation is constitutive of disambiguation, particularly as relates to data that moves across contexts.

Findings reveal that contextualisation in the subsurface domain is largely embedded in domain-specific applications and in the tacit judgement of the practitioners who use them, and fragments when data moves between domains, applications, or out of the source environment. The transition to OSDU reorganises this work rather than removing it, and established applications and workflows coexist with the new. Some context is encoded explicitly through platform mechanisms such as shared data models, Activity Templates, Technical Assurance, and lineage, while the rest continues to live with the people who do the interpretive work. Across this trajectory, data is first disambiguated through categorisation, then made more ambiguous through detachment from its source, then disambiguated a second time through reconstruction in a new

setting. The same trajectory shapes the conditions under which AI can be used on industrial data, since the work of making data legible to AI shifts upstream into the preparation of the data themselves.

The study contributes to research on data work and AI in industrial domains. It offers a process model of how interpretive industrial data are disambiguated through contextualisation as they move across contexts, and extends existing work by showing that disambiguation is non-linear, with a detour through ambiguity between two phases of disambiguating work. It contributes to the AI-at-work literature by locating contextualisation on the data side, rather than only at the point where AI outputs are consumed, and by showing that the transition to a shared data platform reorganises interpretive labour rather than eliminating it.

7.1 Limitations and Future Research

While several measures have been taken to ensure the quality of this thesis, it nonetheless has limitations that should be acknowledged, alongside opportunities for future research.

Firstly, this study is based on a single-case design within the subsurface domain of the oil and gas industry, with a major Norwegian oil and gas operator as the case company. Considering the single-case design and the scale and international presence of the organisation, the case can be considered extreme and thus limits the generalisability of our findings. In addition, the role of contextualisation identified is closely tied to the characteristics of subsurface data, including their indirect and uncertain relationship to physical phenomena (Mikalsen & Monteiro, 2021). Future research could extend this work by conducting multi-case studies across different industrial settings, such as manufacturing or healthcare, to examine whether similar aspects of contextualisation are used to disambiguate data in other domains. This would help clarify which aspects of contextualisation are domain-specific and which reflect more general patterns of data work.

Secondly, a central limitation has been the limited time span of our study. As our study has been ongoing for only 17 weeks, our empirical material captures only a snapshot of ongoing practices. Prior research shows that data and their meanings evolve through continuous interpretation and reuse (Parmiggiani et al., 2024). Longitudinal studies would therefore be valuable to examine how contextualisation processes develop across longer time horizons, especially how they evolve as data is moved to a fully adopted shared data platform.

Thirdly, the study is conducted during an ongoing transition to a shared data platform, which makes it difficult to assess its full implications. Many findings relate to the anticipated capabilities of the platform rather than fully realised practices. In addition, to our knowledge, the adoption of the platform remains limited. Future research could thus follow the long-term effects of such platforms, for example, through ethnographic studies, to examine how contextualisation is reorganised once the platform is more fully embedded in everyday workflows. Regarding adoption, it would also be interesting to study how practitioners engage with and incorporate such platforms into their work, especially when established practices persist alongside the transition of such shared data platforms.

A further limitation concerns the boundary of the process model itself. The model is scoped to interpretive industrial data without a directly observable referent, and we have only demonstrated it in one such domain. Whether the detach-to-reconstruct structure holds where data has a stable, directly accessible referent — or in interpretive domains organised differently from subsurface work — remains an open empirical question rather than an established claim. Future research could test the model in contrasting domains to determine which features of the trajectory are general and which are specific to data whose referent is inaccessible.

Finally, we want to highlight an interesting research direction that revealed itself during our study. This study focuses primarily on contextualisation within a single organisational setting, leaving the broader platform ecosystem out of scope. However, shared data platforms such as OSDU depend on coordination among multiple actors, where data must be standardised and made reusable across

organisational boundaries. Hence, future research could investigate how different actors negotiate and collaborate to establish shared data standards and how contextualisation is negotiated at the ecosystem level.

Bibliography

Aaltonen, A., Alaimo, C., & Kallinikos, J. (2021). The Making of Data Commodities: Data Analytics as an Embedded Process. *Journal of Management Information Systems*, *38*(2), 401–429. https://doi.org/10.1080/07421222.2021.1912928

Aaltonen, A., Alaimo, C., Parmiggiani, E., Stelmaszak, M., Jarvenpaa, S., Kallinikos, J., & Monteiro, E. (2023). What is Missing from Research on Data in Information Systems? Insights from the Inaugural Workshop on Data Research. *Communications of the Association for Information Systems*, *53*(1), 475–490. https://doi.org/10.17705/1CAIS.05320

Ackoff, R. (1989). From Data to Wisdom. *16*(3), 3–9.

Alaimo, C., & Kallinikos, J. (2022). Organizations Decentered: Data Objects, Technology and Knowledge. *Organization Science*, *33*(1), 19–37. https://doi.org/10.1287/orsc.2021.1552

Alaimo, C., Kallinikos, J., & Aaltonen, A. (2020). Data and Value. In *Handbook of Digital Innovation* (pp. 162–178). Edward Elgar Publishing. https://doi.org/10.4337/9781788119986.00022

Bailey, D. E., Leonardi, P. M., & Barley, S. R. (2012). The Lure of the Virtual. *Organization Science*, *23*(5), 1485–1504. https://doi.org/10.1287/orsc.1110.0703

Bøe, A. E. (2015, September). *Diskos 20 år: I oljegeologiens tjeneste*. Oljedirektoratet. https://www.sodir.no/48d68e/globalassets/3-diskos/documents/diskos_jubileumshefte_liten.pdf

Bowen, G. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Burclaff, N. (2025). Research Guides: Oil and Gas Industry: A Research Guide: Introduction. Retrieved February 19, 2026, from https://guides.loc.gov/oil-and-gas-industry/introduction

Davenport, T. H. (2018). From analytics to artificial intelligence. *Journal of Business Analytics*, *1*(2), 73–80. https://doi.org/10.1080/2573234X.2018.1543535

Dubois, A., & Gadde, L.-E. (2002). Systematic combining: An abductive approach to case research. *Journal of Business Research*, *55*(7), 553–560. https://doi.org/10.1016/S0148-2963(00)00195-8

Economist, T. (2017). The world's most valuable resource is no longer oil, but data. *The Economist*. Retrieved May 16, 2026, from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Equinor. (n.d.-a). The digital energy company. Retrieved November 18, 2025, from https://www.equinor.com/energy/digitalisation#data

Equinor. (n.d.-b). Subsurface data and AI in Equinor. Retrieved November 18, 2025, from https://www.equinor.com/energy/data-in-equinor

Felin, T., & Holweg, M. (2024). Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, *9*(4), 346–371. https://doi.org/10.1287/stsc.2024.0189

Fischer, H., Wiener, M., Strahringer, S., Kotlarsky, J., & Bley, K. (2023). Data-Driven Organizations: Review, Conceptual Framework, and Empirical Illustration. *Australasian Journal of Information Systems*, *27*. https://doi.org/10.3127/ajis.v27i0.4425

Gerring, J. (2004). What Is a Case Study and What Is It Good for? *American Political Science Review*, *98*(2), 341–354. https://doi.org/10.1017/S0003055404001182

Gibbert, M., Ruigrok, W., & Wicki, B. (2008). What Passes as a Rigorous Case Study? *Strategic Management Journal*, *29*. https://doi.org/10.1002/smj.722

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational Research Methods*, *16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Gitelman, L. (2013). *"Raw data" is an oxymoron*. MIT press.

Havindustritilsynet. (2024a, April). *Sikkerhet og ansvar - Forstå det norske regimet*. https://www.havtil.no/contentassets/2d8521c6cb704e71ab71a7a975444c46/sikkerhet-og-ansvar_-forsta-det-norske-regimet.pdf

Havindustritilsynet. (2024b, July). Hva er ambisjonen for sikkerheten? Retrieved February 25, 2026, from https://www.havtil.no/om-oss/sikkerhet-og-ansvar-forsta-det-norske-regimet/sikkerhet/

IADC. (2012, May). Delineation Well. Retrieved May 12, 2026, from https://iadclexicon.org/delineation-well/

Jarvis, R., & Bøklepp, B. (2026, March). Evolving the trusted data ecosystem for AI, geoscience workflows, and business decisions [Powerpoint presentation].

Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems*, *28*(1), 3–16. https://doi.org/10.1016/j.jsis.2018.10.005

Kitchin, R. (2022). *The Data Revolution: A Critical Analysis of Big Data, Open data & Data Infrastructures* (Second edition). Sage.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Quarterly*, *45*(3), 1501–1526. https://doi.org/10.25300/MISQ/2021/16564

Markussen, C., van der Meulen, M., van de Merwe, K., & Kvinnesland, K. (2024, December). *Kunnskapsoversikt knyttet til forsvarlig bruk av kunstig intelligens i petroleumssektoren* (tech. rep. No. 2024-1519). DNV. https://www.havtil.no/contentassets/ef58508a2e4641aebba4091811795020/dnv-

2024-1519-kunnskapsoversikt-forsvarlig-bruk-ki-petroleumssektoren-rev-1.pdf

Maxwell, J., & Miller, B. (2008). Categorizing and connecting strategies in qualitative data analysis. *Handbook of Emergent Methods*, 461–477.

Mikalsen, M., & Monteiro, E. (2021). Acting with Inherently Uncertain Data: Practices of Data-Centric Knowing. *Journal of the Association for Information Systems*, *22*(6), 1715–1735. https://doi.org/10.17705/1jais.00722

Monteiro, E. (2022). *Digital Oil: Machineries of Knowing*. The MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/14604.001.0001

Myers, M. D. (1997). Qualitative Research in Information Systems. *MISQ Discovery*. https://doi.org/10.2307/249422

NorskPetroleum. (n.d.). Lisenser (utvinningstillatelser). Retrieved February 23, 2026, from https://www.norskpetroleum.no/fakta/lisenser/

NorskPetroleum. (2025, April). The Petroleum Act and the licensing system. Retrieved January 28, 2026, from https://www.norskpetroleum.no/en/framework/the-petroleum-act-and-the-licensing-system/

NorskPetroleum. (2026, December). The government's revenues. Retrieved February 23, 2026, from https://www.norskpetroleum.no/en/economy/governments-revenues/

OSDU. (2024, August). Guides/Chapters/06-LifecycleProperties.md · master · OSDU / OSDU Data Definitions / Data Definitions · GitLab. Retrieved May 16, 2026, from https://community.opengroup.org/osdu/data/data-definitions/-/blob/master/Guides/Chapters/06-LifecycleProperties.md?ref_type=heads

Østerlie, T., & Monteiro, E. (2020). Digital sand: The becoming of digital representations. *Information and Organization*, *30*(1), 100275. https://doi.org/10.1016/j.infoandorg.2019.100275

Parmiggiani, E., Amagyei, N. K., & Kollerud, S. K. S. (2024). Data curation as anticipatory generification in data infrastructure. *European Journal of*

Information Systems, *33*(5), 748–767. https://doi.org/10.1080/0960085X.2023.2232333

Parmiggiani, E., Østerlie, T., & Almklov, P. G. (2022). In the Backrooms of Data Science. *Journal of the Association for Information Systems*, *23*(1), 139–164. https://doi.org/10.17705/1jais.00718

Peirce, C. S., Hartshorne, C., Weiss, P., & Burks, A. W. (1931). *Collected Papers of Charles Sanders Peirce* [Issue: v. 2]. Harvard University Press. https://books.google.no/books?id=B9jWAAAAMAAJ

Recker, J., Indulska, M., Green, P., Burton-Jones, A., & Weber, R. (2019). Information Systems as Representations: A Review of the Theory and Evidence. *Journal of the Association for Information Systems*, *20*(6), 735–786. https://doi.org/10.17705/1jais.00550

Smith-Solbakken, M., & Dahle, E. A. (2025, November). Alexander Kielland-ulykken. Retrieved February 25, 2026, from https://snl.no/Alexander_Kielland-ulykken

Sokkeldirektoratet. (n.d.). Guidelines. Retrieved May 12, 2026, from https://www.sodir.no/en/regulations/guidelines/

Sokkeldirektoratet. (2024, July). Release of data. Retrieved May 19, 2026, from https://www.sodir.no/en/facts/data-and-analyses/release-of-data/

Tollaksen, T. G., Ryggvik, H., & Smith-Solbakken, M. (2025, September). Equinor SNL. Retrieved October 28, 2025, from https://snl.no/Equinor

Tuomi, I. (1999). Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. *Journal of Management Information Systems*, *16*(3), 103–117. https://doi.org/10.1080/07421222.1999.11518258

van den Broek, E. (2025). Unpacking AI at work: Data work, knowledge work, and values work. *Information and Organization*, *35*(3), 100584. https://doi.org/10.1016/j.infoandorg.2025.100584

van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly*, *45*(3), 1557–1580. https://doi.org/10.25300/MISQ/2021/16559

Waardenburg, L., & Sergeeva, A. (2021). In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. *Organization Science*, *33*. https://doi.org/10.1287/orsc.2021.1544

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems*, *4*(2), 74–81. https://doi.org/10.1057/ejis.1995.9

Xu, D., Stelmaszak, M., & Aaltonen, A. (2025). What is Changing the Game in Data Research? Insights from the "Innovating in Data-based Reality" Professional Development Workshop. *Communications of the Association for Information Systems*, *56*(1), 194–208. https://doi.org/10.17705/1CAIS.05608

Yin, R. K. (2018). *Case study research: Design and methods* (6. edition). SAGE.

Appendix A – Interview Guides

A.1 Interview Guide 1

Introduction

- Introduction and study purpose
- Consent and confidentiality assurance

Background & Context

- Can you describe your current role and main responsibilities, and how long you've been in your position at Equinor?
- What does a typical week look like for you, particularly in terms of data-related activities?
  - Where in the subsurface workflow do you primarily work?
- What decisions do you influence or make yourself?

The Static Reservoir Model

- Can you take us through, in broad terms, how a static reservoir model is constructed?
- How do seismic and well data interact during this process?
- What parts of the model are considered facts vs. interpretations?
- Where do you see limitations or ambiguities in the model?

Data Quality, Uncertainty, & Trust

- What types of subsurface data do you work with most (seismic, well, models, others)?
  - Where do you experience the greatest uncertainty in this data?

- How is data quality assessed for subsurface data in your area?
- How is uncertainty typically communicated in your work and to management or decision-makers?
- How do you usually validate interpretations or results?
- How do disagreements between experts or colleagues get resolved?
  - When people disagree, is it usually because of the data itself or because it is interpreted differently?

Data Foundations

- How important is it for you to know where the data comes from and how it has been processed?
  - What happens in practice when that information about data origin or processing is missing or unclear?
- How important are standards and data definitions in seismic interpretations and reservoir modeling?
  - Where do they help?
  - Where do they oversimplify or constrain interpretations?
- Have you experienced situations where standards or data definitions made data easier to reuse, but harder to interpret correctly?

Subsurface Data, AI, & ML

- How is AI/ML currently used in your area of subsurface work?
- What are the main challenges in making subsurface data "AI-ready"?
  - Follow-up: Issues with data quality, standardization, accessibility, representing uncertainty?
- Can you provide examples of how improved data governance has enabled or could enable AI/automation in seismic interpretation and reservoir modeling?
  - Where do you see real value?
  - Where are the limits?
- Who decides whether subsurface data is "good enough" to be used for modeling or automation?

Challenges & Future Vision

- Are there specific areas where better data governance could immediately impact AI/automation capabilities?
- Where do you see the biggest gaps today between how data is produced, interpreted, and reused?
- Looking ahead, what would need to change for AI to play a more meaningful and enabling role in subsurface decision-making/workflow?

Closing

- Is there anything important about data, AI, or oil and gas exploration that we haven't discussed?
- Could you recommend other colleagues who might provide valuable perspectives on these topics?
- Would you be available for follow-up questions if needed?

A.2 Interview Guide 2

Mekanismer som reduserer tvetydighet i brønnloggdata

- Hva gjøres for å gjøre data mindre tvetydige?
- Når dere veksler mellom ulike datatyper, hvordan sørger dere for å bevare innholdet eller meningen i denne overgangen?
- Hva er det som helt konkret går tapt når data flyttes ut av applikasjoner som Petrel?
- Hva gjøres for å bevare eller gjeninnføre det som går tapt, for eksempel i OSDU?
- Hvordan kommuniserer dere videre mening eller tolkning til personer lengre ned i verdikjeden, for eksempel personer uten domenekunnskap?
- Hvilken type kontekstuell informasjon er nødvendig for å tolke brønnloggdata korrekt?
- Hvordan dokumenteres eller kommuniseres vanligvis tolkninger og antakelser som ligger til grunn for analysen?
- Finnes det bestemte metadatafelt eller datastrukturer som er spesielt viktige for å bevare meningen i brønnloggdata?
- Hvilken rolle spiller domenekunnskap i tolkningen av målingene?

OSDU og endringer i hvordan data representeres

- Hvordan forsøker OSDUs datamodell for brønnloggdata å representere både selve dataene og konteksten rundt dem?
- Hvor viktig er det å fange den bredere konteksten rundt dataene?
  - For eksempel formålet med målingene, hvilket prosjekt de ble produsert i, og hvem som produserte eller tolket dem.
- Hvilke typer informasjon som tidligere lå inne i applikasjoner blir nå fanget opp i datamodellen?
- Hva blir fortsatt ikke fanget opp?
- Har OSDU endret hvordan tolkningens linjeføring (lineage) eller opphav (provenance) dokumenteres?
- Hva er rollen til Activity Templates / Activity Models?

- I praksis, bidrar OSDU til å redusere tvetydighet når subsurfacedata deles på tvers av fagdisipliner?
- Kan du forklare mer detaljert hvordan Activity Templates / Activity Models fungerer for brønnloggdata?
  - Hva er hovedformålet med disse?
  - Er det primært ment for å kontekstualisere dataene — altså hvor de kommer fra, hvordan de er produsert, og hvem som har produsert dem?

Gjenstående utfordringer

- Finnes det aspekter ved tolkning av brønnloggdata som fortsatt er vanskelige å representere i standardiserte datamodeller?
- Finnes det typer kontekstuell kunnskap som er vanskelige å formalisere?
- Hva er de viktigste hindringene for å gjøre subsurface-data mer forståelige og tolkningsbare på tvers av systemer?

Datagrunnlag for AI

- Hvordan ser du på muligheten for å lage "foundation models" før og etter OSDU-initiativet?
- Hvilke utfordringer ser du i forberedelsen av subsurface-data for å lage "foundation models"?
- Finnes det spesifikke typer tvetydighet i subsurface-data som gjør AI-utvikling vanskelig?
  - Manglende data lineage og opphav?
  - Ulike datatyper som må avstemmes?

Appendix B – Consent Form

Vil du delta i forskningsprosjektet om subsurface-data og kunstig intelligens i Equinor?

Formålet med prosjektet
Dette er en forespørsel til deg om du vil delta i et forskningsprosjekt. Forskningsprosjektet har som mål å undersøke hvilke forutsetninger som må være til stede for at subsurface-data kan tas i bruk i maskinlærings- og KI-løsninger. Spesielt fokuseres det på utfordringer knyttet til data som er utviklet for menneskelig tolkning, slik som reservoarmodeller, seismiske tolkninger og brønndata.

Formålet med prosjektet er å levere en avsluttende masteroppgave ved Universitetet i Oslo knyttet til programmet Informatikk: Digital Økonomi og Ledelse våren 2026.

Hvorfor får du spørsmål om å delta?
Du får denne forespørselen fordi du er vurdert som en relevant informant knyttet til studiet, enten som en leder- eller mellomleder i Equinor, eller som en ekspert på fagområdet Data Management, Data Governance, Data Engineering, Geologi, Geofysikk eller lignende.

Dette vil være en kvalitativ studie med omtrent 8-12 intervjuer.

Hvem er ansvarlig for forskningsprosjektet?
Universitetet i Oslo: Institutt for Informatikk er ansvarlig for personopplysningene som behandles i prosjektet.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Hva innebærer det for deg å delta?
Hovedmetoden for innsamling av data vil være intervju med elektroniske notater. Ved tillatelse fra intervjuobjektet vil det bli tatt lydopptak med UiO diktafon eller video- og

lydopptak med UiO Teams for transkribering. Intervjuene vil være på omtrent en time og det er ønskelig med flere intervjuer om ulike temaer dersom det ses som aktuelt.

Navn, kontaktopplysninger og beskrivelse av arbeidserfaring eller stilling vil samles inn. Navn og kontaktopplysninger vil anonymiseres ved publisering av oppgaven.

Kort om personvern
Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler personopplysningene konfidensielt og i samsvar med personvernregelverket. Du kan lese mer om personvern på neste side.

Med vennlig hilsen

Dragana Paparova
(Forsker/veileder)

Stefan Spanic & Eirik Borgen Egge
Masterstudenter

Personvern

Utdypende om personvern – hvordan vi oppbevarer og bruker dine opplysninger
Ingen sensitive personopplysninger (jf. Personvernforordningens artikkel 9 og 10) vil bli samlet. Personlige opplysninger om deg vil kun benyttes til formålene beskrevet i dette informasjonsskrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

Dersom det kommer frem sensitive personopplysninger i intervjuet, vil disse bli fjernet umiddelbart samme dag. Det er kun prosjektgruppen (studentene som gjennomfører prosjektet) og veileder som vil ha tilgang til dataen, og det som oppbevares av anonymisert rapportering fra intervjuet vil følge Universitetet i Oslo sine rutiner for sikker oppbevaring.

Din kontaktinformasjon vil kun være kjent for gruppemedlemmene. Data kan ettersendes deg ved ønske.

Hva gir oss rett til å behandle personopplysninger om deg?
Vi behandler opplysninger om deg basert på ditt samtykke. Så lenge du kan identifiseres i datamaterialet, har du rett til:

Dine rettigheter
- å be om innsyn i hvilke opplysninger vi behandler om deg, og få utlevert en kopi av opplysningene,
- å få rettet opplysninger om deg som er feil eller misvisende,
- å få slettet personopplysninger om deg,
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger.

Vi vil gi deg en begrunnelse hvis vi mener at du ikke kan identifiseres, eller at rettighetene ikke kan utøves.

Hva skjer med personopplysningene dine når forskningsprosjektet avsluttes?
Alle opptak og eventuelle notater fra intervjuet blir slettet senest 20.06.2026. Informasjonen som kommer fram i intervjuet vil kun brukes i masteroppgaven og prosjektrapporten. Prosjektrapporten vil ikke bli publisert og vil kun bli lest av gruppemedlemmene, veileder og sensor i faget IN5560 Data Governance. Masteroppgaven vil bli publisert til Universitetet i Oslo sitt vitenarkiv DUO. Alle sitater og beskrivelser vil bli anonymisert. Det vil ikke publiseres noen personopplysninger.

Spørsmål
Hvis du har spørsmål angående intervjuet, eller ønsker å benytte deg av dine rettigheter, ta kontakt med Eirik / Stefan på e-post: eirikbeg@uio.no / stefansp@uio.no

Før intervjuet begynner, ber vi deg om å samtykke i deltagelsen ved å undertegne på at du har lest og forstått informasjonen på dette arket, og ønsker å stille opp til lydintervju.

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om forskningsprosjektet om subsurface-data og kunstig intelligens i Equinor, og har fått anledning til å stille spørsmål. Jeg samtykker til:

- ☐ å delta i intervju
- ☐ at opplysninger om meg publiseres slik at jeg kan gjenkjennes gjennom nåværende og tidligere arbeidsforhold.
- ☐ at mine personopplysninger lagres etter prosjektslutt i anonymisert format.

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

---

(Signert av prosjektdeltaker, dato)