Data sourcing in the context of software product innovation

Tor Sporsem
NTNU
Trondheim, Norway
tor.sporsem@sintef.no

Anastasiia Tkalich
NTNU
Trondheim, Norway
anastasiia.tkalich@sintef.no

***Abstract*—Companies are increasingly relying on data to provide business offerings; however, data sourcing can be problematic when organizations collaborate across the borders. At the same time, data collaborations are crucial to succeed in software product innovation. This study seeks to address this problem by examining how companies go about data sourcing in the context of software product innovation. We report preliminary findings from a longitudinal case study of a maritime company. The results highlight the themes such as *Data as an asset for software product innovation*, *Using data for automation*, *Data partnership and ownership*, and *Data sourcing as a novel discipline*. Comparing our findings with earlier literature, we concluded that actors could have their implicit views on data, however not be aware of different modes of data governance. Data sharing was problematic when data partnership practices were not in place. The results are useful for companies who develop software products based on data and other companies that potentially encounter problems with data partnerships.**

Keywords: data sourcing, software product innovation, data partnership

I. Introduction

This study is motivated by the need for more knowledge on data acquisition and especially the role of data acquisition in software product innovation. Companies such as Facebook increasingly rely on data as a foundation for their businesses. Data has become a valuable asset that can be exchanged for other services [1]. However, the problem of data acquisition has not been considered in the Information systems (IS) sourcing research until recently [2]. At the same time, today's software product innovation often relies on data to provide the new business offering. Our partner Marcomp (real name suppressed for anonymity) is an example of this with a portfolio of emerging software products for data-driven maritime inspection. In this example, we see that data sourcing appears to be challenging and that there is little awareness around this issue in the innovation process. The objective of this study is, therefore, twofold: 1) to contribute to more knowledge on data sourcing, and 2) to contribute to solving the practical problem of data sourcing in software product innovation.

Following the definition of Jarvenpaa and Markus [2], we refer to *data sourcing* as procuring, licensing, and accessing data from an internal or external entity (supplier), which constitutes inter-organizational relationships. *Software product innovation* can be defined as the creation and introduction of a novel software product to the market [3]. An emerging problem is accessing data from third parties to conduct software product innovations. This is the problem that we are observing in our case company. The purpose of this research is thus obtaining knowledge on how Marcomp's employees gain access to data today and suggest new ways for the future. To specify, these methods can, i.e., be contracts, tacit agreements, etc., not technical methods of accessing data. Without this knowledge, Marcomp will not be able to offer new innovative software products and may lose competitive advantage. Our motivation for this study is to create new knowledge and insight that can aid our partner and the entire maritime industry. Other beneficiaries of this research result can be companies that develop software products based on data and other companies that potentially encounter problems with data sourcing. With that in mind, our research question is the following:

RQ: How does a maritime company go about data sourcing in the context of software product innovation?

The paper will contribute to the scope of EReMCIS as it describes the implementation of new digital systems in a private organization (Theme 1). The results are also shedding light on a method of accessing data and evaluate this method, thus fitting the formal requirement nr.1 for an EReMCIS paper.

In the remainder of this paper, we describe the earlier research and present a gap in knowledge in chapter II, followed by a description of research methods in chapter III, including data collection and analysis. Chapter IV presents our findings before they are discussed in relation to earlier research in chapter V. The paper ends with a conclusion, proposal of future research, and the study's limitations in chapter VI.

II. Background

Data sourcing is becoming a precondition for innovation [4]. Digital innovation nowadays often involves AI/machine learning and analytics that rely on data to function as intended. What types of data are necessary, what quality and what format, is often a well-discussed topic in software development. However, the methods to *access* data are often forgotten. As noted by Constantiou and Kallinikos [5] "algorithms without data are just a mathematical fiction". At the same time, several challenges exist with regard to sharing data across organizations, which is primarily due to commercial sensitivity of data and privacy risks involved [4]. To address these challenges, the researchers have explored various views on data, different modes of data partnership and practices, and that allow organizations collaborate on data sharing.

Jarvenpaa and Markus [2] summarized the implicit views on data from the literature: *commodity view*, *process view* and *relational view*. It is important to understand these views to better understand which premises data partnerships are based on. According to the authors, the *commodity view* prevails in the IS research. Data is seen as a resource for a final good or service just like any other resource. An example of commodity view is open pool exchange in academic research when researchers publicly deposit their data. The *process view* focuses on the value that data provides through its use in operations. The value of the data can increase as a

consequence of its transformation and recombination. Example of the process view on data is Multicenter AIDS Cohort Study [6], where the participating organizations agreed to standardize data across all samples. The *relational view* is based on trusting relationships between organizations where the value of data is co-generated. An example of such view is Structural Genomic Consortium [7], where industry and academia collaborated to discover new drugs. Even though Jarvenpaa and Markus [2] describe the three views of data, the examples they provide originated primarily from the domain and research, and not from the domain of commercial software product innovation.

In the commercial domain, Broek and Veenstra [4] described four modes of data governance that allow organizations to collaborate on data sharing. Exploring how companies can govern their data relationships is one way to solve the problem of data acquisition from external suppliers, as in our case. The authors describe three modes of governance that they derive from empirical cases: *Bazaar*, *Hierarchy* and *Network.* These modes differ in who controls the data. In the *Bazaar* mode the data is openly available and thus is controlled by anyone. In the *Hierarchical* mode the data is controlled by a dominant organization (e.g. an insurance company). The *Network* mode is characterized by mutual sharing of data between the members of the network, which is based on mutual trust. In this mode individual organizations remain in control of their respective data but share it with the other members through specific mechanisms. Surprisingly, no organizations in this study were buying or selling through regular contracts.

To facilitated data partnership, four practices were suggested to increase the return from data and analytics: (1) identifying/evaluating new partners, (2) incentivizing organizations to partner, (3) making explicit how value will be shared, and (4) making explicit how partnership conflicts will be communicated and resolved [8].

III. METHODS

Case study was chosen as the research strategy as described by Yin [9], because we wanted an in-depth understanding how actual people gain access to data in their existing context. Understanding their complex context and relationships between them are important to really comprehend the practical problem and define a research agenda. Semi-structured interviews and document analysis from Marcomp were our data generating methods, as they are described in [10]. Interviews can provide insight into respondents' perspectives of accessing data and are easily conducted through virtual communication and can be constructed as open and explorative. Documents from Marcomp allowed us to better understand the company's context. The transcribed interviews and documents were analyzed qualitatively through inductive coding using Nvivo to reveal different themes and concepts. In this way we were hoping to connect our inductively discovered concepts to exciting theory and identify a research gap.

There were two participants from Marcomp in our study both leading novel software product ideas (product managers). We interviewed each of them twice.

When it comes to our research paradigm, we took a standpoint that different truths of what data mean are held by different respondents. They will probably fill the term with different meanings and interpret the concept differently. For instance, can the same data hold different value for different stakeholders or serve different purposes. This points us in the direction of interpretivism [10]. At the same time, one can argue that data is a hard object that does not change as to what point of view one takes, which points us in a positivistic direction. Since our study focuses on how different respondents' perspectives and relationships affect data access, we end up leaning towards interpretivism. Additionally, we as researchers are affected by earlier work and our partners' interests and focus. How we interpret our findings will be influenced by our point of departure and our worldview.

Data analysis was performed in two steps, as described by Oates [10]. First, textual data was entered into the qualitative data analysis tool NVivo. We coded the data inductively, which means that codes emerge from our data, and 23 codes were created. Then we categorized our codes into four themes, keeping our selected literature in mind and thus following the deductive approach (Table I). The four themes make up the structure of the following chapter on findings.

TABLE I. ILLUSTRATION OF THE CODING PROCESS

Data instance	***Inductive code***	***Theme***
"When we initially approached them, we told them about our purpose, and they were like – yeah, win-win."	Interaction with the third-party vendor	Data partnership and ownership
"Everyone sees value in sensor data. Everyone is trying to find a business model. … Everyone knows that there is a lot to be learned, and many businesses to be developed"	Sensor data is valuable for developing new business	Data as an asset for software product innovation

IV. FINDINGS

In this chapter we describe data sourcing challenges met by two initiatives of software product innovation. The findings are structured according to the themes we have identified through the data analysis and thus written presented chronologically as they are dependent on each other.

The goal of the software product innovation was to replace rule-based and manual inspection processes with risk-based, targeted and data-driven by using various sources of ship data. In such way today's manual, prescriptive, calendar-based inspection and maintenance regime can be more accurate, efficient and simple.

A. Data as an asset for software product innovation

Since the product ideas were relying on data, data became an asset that the new offerings were relying on. The product ideas of our informants were based on using different sources of sensor data to provide real-life insight into vessels' conditions, such as carbon emissions (product idea 1) and ballast water (product idea 2). One product manager emphasized that the sensor data has a clear business value: "*Everyone sees value in sensor data. Everyone is trying to find a business model. … Everyone knows that there is a lot to be learned, and many businesses to be developed*". Instead of inspectors being on the ships and reading data from instruments, then manually entering it into reports, the new idea was collected centrally, combined, analyzed and visualized to provide insight both for the vessel-owners and the inspectors. Some data sources were easily available, whereas others were not. For example, product idea 1 was relying on vessel data that were a property of Marcomp

combined with publicly available (e.g. ships cargo data), whereas product idea 2 was based on data from sensors that were owned by a third party.

B. Using data for automation

Product idea 2 was to automatically collect sensor data from ships and compare it to requirements set by the harbor, proving that a ship is in line. Today, this job is done by manually inserting a USB stick into the ship machinery, downloading the sensor data and e-mailing it to Marcomp where a human assesses the data and produces a certificate. All this would be done automatically. Luckily for Marcomp, the vendor of the ship machinery with the sensor data offered a cloud service to upload sensor data automatically as an option for shipowners.

C. Data partnership and ownership

Marcomp's product manager 2 met with the vendor to discuss how they could gain access to their cloud data and agreed that sharing data would be beneficial for both parties. The product manager said: *"When we initially approached them, we told them about our purpose, and they were like – yeah, win-win."* The product manager had several meetings with the vendor to secure collaboration and set up an experiment on extracting data from the cloud service. An API was suggested put up to gain automatic access to the cloud data.

However, the vendor started exploring possibilities in making similar business models themselves, realizing the sensor data could be valuable. "*They started asking us what was our business case here? Why were we doing this? Why can't you use data you collect on the ships?"* the product manager recounted. The vendor eventually became reluctant to share their cloud data.

Marcomp then engaged shipowners to try and find a solution. The product manager said: "*Shipowners are more than willing to share their data, while the device vendor is very much aware of the data's value.*" All agreed that the shipowners who operated the installed sensors also owned the sensor data. But the vendor concluded that if the data was uploaded to the vendor's cloud service they defined it as "derived data," hence becoming their property. Marcomp's legal team together with management, is still trying to find a solution for accessing the cloud data. The negotiations have been ongoing for more than half a year.

D. Data sourcing is a novel discipline

Using automatically gathered sensor data was a new idea in Marcomp and this effort of finding a solution for sharing data was ground-breaking for the company. One product manager said: "*Data transfer is new for everyone, right?*" This was the first time both Marcomp and the vendor is handling a case like this. The product manager told us that the third-party vendor is searching for how they can make value from data and find new business models, ahead of Marcomp: "*They are really testing if they can do that service [of making compliance certificates] instead of us.*"

V. Discussion

Earlier research has identified that sharing data across organizations can be problematic [4]. At the same time, data sourcing is a precondition for innovation, as is also the case in the company we have studied. We thus asked a research question *How do companies go about data sourcing in the context of software product innovation?* To answer the research question, we report findings from a case study at a maritime company. We have reported four themes that reflect various aspects of data sourcing: *Data as an asset for software product innovation*, *Using data for automation*, *Data partnership and ownership* and *Data sourcing as a novel discipline*. The findings indicate that data (and specifically sensor data in our context) have a high business value because they allow to create novel software products. The products serve the goal of automation and simplification of the existing analogue processes (such as maintenance and validating compliance with the regulations). In this context data partnerships play an important role because the data needed for creating the new software products can be distributed across different actors. As our findings show, the actors can compete for the same data. In such situations, data partnerships become central as they facilitate sharing of data value among different firms.

We will draw on earlier research on data sharing to discuss our finding in light of *implicit view on data* [2], *modes of data governance* [4] and *data partnership practices* [8]

A. Implicit views on data

Earlier research has formulated three implicit views on data that actors can hold when sharing data: commodity view, process view and relationship view [2]. Our case illustrated that software product managers also have their implicit assumptions about data prior to committing to data partnership. Both product managers' view can be compared to the *process view* as they were seeking to leverage the data value through operation towards their customer. This contributes to empirical findings from the world of business instead of the world of research as offered by Jarvenpaa and Markus [2]

B. Governance modes

Broek and Veenstra [4] described four modes of data governance that allow organizations to collaborate on data sharing. In our case we have observed that actors may not be aware of which governance mode is dominating in their relationships with the In our case we have observed that Broek and Veenstra [4] described four modes of data governance that allow organizations to collaborate on data sharing. In our case we have observed that actors may not be aware of which governance mode is dominating in their relationships with the third parties. For example, product manager 2 was initially expecting mutual sharing of data between the members of the network, which is based on mutual trust (*Network* governance mode). She based her data sourcing strategy on tacit agreement and personal relationships with the third party. However, it turned out that the relationships with the third party owner of the required data were hierarchical, meaning that the data turned out to be controlled by the third party. Commercial sensitivity of data was an obstacle with regard to data sharing [4]. This was an obstacle in the innovation process. On the other hand, this was not an obstacle for the other product that did not have to rely on data owned by third-parties. Some of the necessary data was openly available and thus reminded the *Bazaar* mode of governance. Our findings are in accordance with Broek and Veenstra [4] in showing that traditional contracts for buying and selling data were not used to control partnerships.

C. Data partnership practices

Several practices have been suggested in order to facilitate data partnership. Our case illustrates how the product managers were not initially aware of these practices, which led to complications with data sourcing. Product manager 2 did not identify the third-party vendor as new data partners, did not incentivize it to partner, didn't make it explicit how the value will be shared and how the partnership conflict could be resolved [8].

VI. Conclusions

Data sourcing across organizational borders has been described as problematic. Against this background, this study has examined how product managers in a maritime company go about data sourcing for their novel software products. We have found that one product manager was relying on data that was openly available, whereas the other had to acquire data from the first party. Comparing our findings with earlier literature, we concluded that actors could have their implicit views on data, however not be aware of different modes of data governance. Data sharing was problematic when data partnership practices were not in place.

A. Limitations and future work

The limitation of this study is its timeline, which allowed only for preliminary data collection (4 interviews and documents). In the future, we will continue to expand our data collection within the case company as they continue innovating software products. We urge fellow researchers to look into how companies solve their data sourcing challenges as there exist few empirical findings, and it seems highly relevant to most industries.

VII. References

[1] A. Pentland, "Personal Data: The Emergence of a New Asset Class." http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf.

[2] S. L. Jarvenpaa and M. L. Markus, "Data Sourcing and Data Partnerships: Opportunities for IS Sourcing Research," in *Information Systems Outsourcing: The Era of Digital Transformation*, R. Hirschheim, A. Heinzl, and J. Dibbern, Eds. Cham: Springer International Publishing, 2020, pp. 61–79.

[3] H. Edison, N. M. Smørsgård, X. Wang, and P. Abrahamsson, "Lean Internal Startups for Software Product Innovation in Large Companies: Enablers and Inhibitors," *Journal of Systems and Software*, vol. 135, pp. 69–87, Jan. 2018, doi: 10.1016/j.jss.2017.09.034.

[4] van den Broek and van Veenstra, "Modes of Governance in Inter-Organizational Data Collaborations," in *ECIS 2015 Completed research paper*, 2015, p. 13.

[5] I. D. Constantiou and J. Kallinikos, "New Games, New Rules: Big Data and the Changing Context of Strategy," *Journal of Information Technology*, vol. 30, no. 1, pp. 44–57, Mar. 2015, doi: 10.1057/jit.2014.17.

[6] D. Ribes and J. B. Polk, "Organizing for ontological change: The kernel of an AIDS research infrastructure," *Soc Stud Sci*, vol. 45, no. 2, pp. 214–241, Apr. 2015, doi: 10.1177/0306312714558136.

[7] M. Perkmann and H. Schildt, "Open data partnerships between firms and universities: The role of boundary organizations," *Research Policy*, vol. 44, no. 5, pp. 1133–1143, Jun. 2015, doi: 10.1016/j.respol.2014.12.006.

[8] Wixom, B, "Data monetization: Generating financial returns from data and analytics— Summary of survey findings," *CISR WP*, no. 427, 2019.

[9] R. Yin, *Case study research: Design and methods*, 4th ed. Sage, 2009.

[10] B. J. Oates, *Researching information systems and computing*. Sage, 2005.