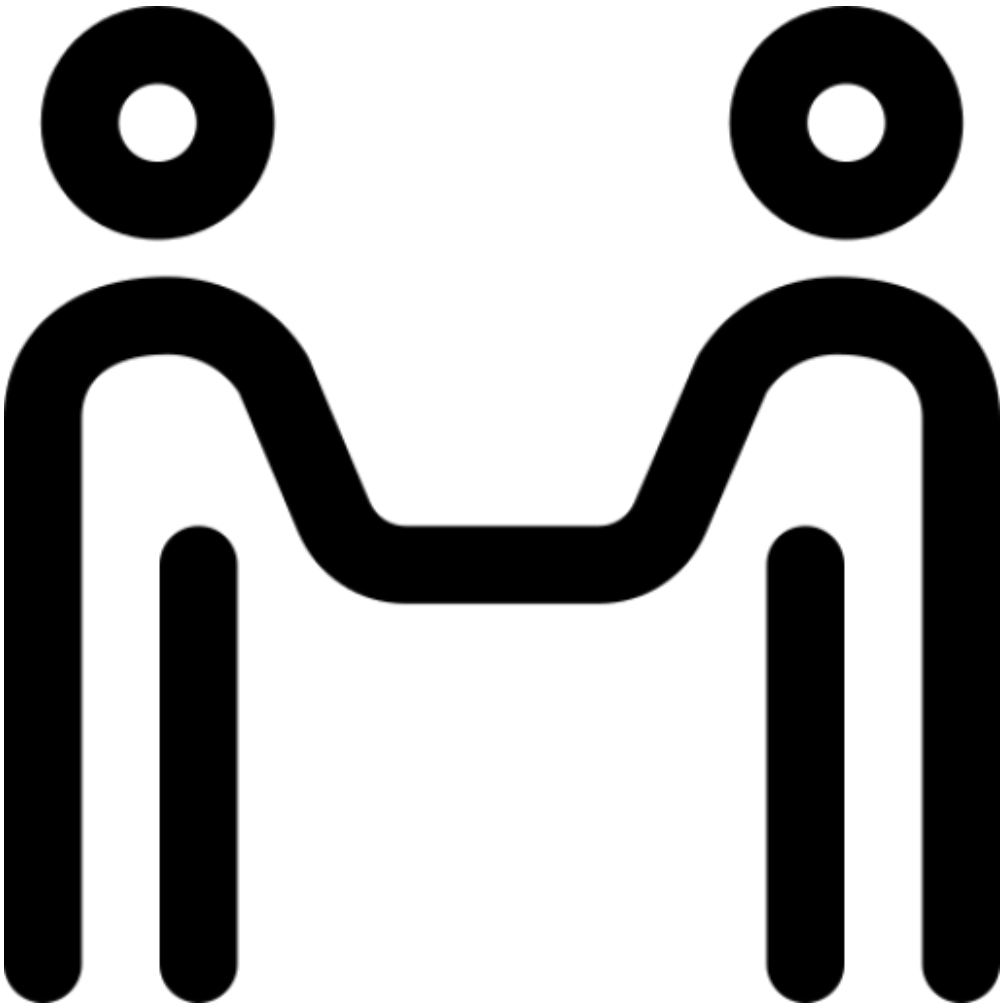
ICT
SICS



Open Collaborative Data – a pre-study on an emerging practice

Thomas Olsson, RISE Research Institutes of Sweden
Per Runeson, Lund University
Sofie Westerdahl, Mobile Heights

RISE Report 2019:77

Open Collaborative Data – a pre-study on an emerging practice

Thomas Olsson, RISE Research Institutes of Sweden
Per Runeson, Lund University
Sofie Westerdahl, Mobile Heights

Abstract

Open Collaborative Data – a pre-study on an emerging practice

Data intense defined software is becoming more and more prevalent, especially with the advent of machine learning and artificial intelligence. With data intense systems comes both challenges – to continue to collect and maintain quality – and opportunities – open innovation by sharing with others.

To understand challenges and opportunities with ODC, we ran 5 focus groups (4 in Lund and 1 in Kista) with companies and public organizations. We had 27 participants from 22 organizations.

Despite an interest to participate and understanding of the potentials of the subject, the overall maturity is low and ODC is rare. For ODC to be successful, there is a need to study technical, organizational, business, and legal aspects further.

Key words:

RISE Research Institutes of Sweden AB

RISE Report 2019:77
ISBN: 978-91-89049-06-2
Lund 2019

Content

Acknowledgements

First and foremost, we would like to thank all the participants of the focus group meetings! We would also like to thank our colleagues who have contributed with valuable feedback to the project proposal and design of the pre-study. Lastly, we would like to thank Vinnova for funding our pre-study.

Summary

This report summarizes the project activities and results from the pre-study on “Open Data Collaboration as an Innovation Platform for Machine Learning”, funded within the Vinnova program for “Ground-breaking Ideas in Industrial Development”.

Development of new products and services depend highly on access to data. “Data is the new oil” is a frequently used slogan. This is particularly true for machine learning and artificial intelligence (ML/AI) applications where amazing progress is made in recent years.

As the development progresses in a field, basic functions will turn into commodity, i.e. features that are offered by most competitors and thus are not contributing to the competitive edge. However, the features are still expected by the users and customers to be part of the products. Thus, commodity features should be delivered at the lowest cost possible and resources should focus on development of value-creating features instead.

We propose the concept of Open Data Collaboration to address this challenge, in line with Open Source Software (OSS) being used to share costs for maintaining platforms and commodity features. Similar to software, data can be either a competitive edge or a commodity. Furthermore, to facilitate open innovation, sharing data with other organizations can be a mechanism. ODC concerns the practices and principles around governance, platforms, processes, etc., to establishing and maintaining collaborations around data.

To explore the practical needs and implications for the concept, we organized a workshop series, where industry, public sector and academia met in focus group meetings to discuss 1) needs for data, 2) sharing data, and 3) enabler and hinders for sharing data. The five focus groups comprised 27 participants from 22 companies and public agencies (primarily municipalities).

The main outcome from the focus groups is that companies and public agencies find the open data collaboration to be an interesting concept, although neither the concept nor the actors are mature enough to put into it practice. We identified the following themes in the study: 1) Open data strategies, 2) Data purchase, 3) Mindset for sharing, 4) Costs for data, 5) Quality and trust, 6) Standards and APIs, and 7) Competence.

Each of these constitute enablers and hinders for ODC, which warrant further studies. We therefore propose further work on technical, organizational, business, and legal issues of open data collaboration.

The project also created a network of interested partners, both participants in the focus groups, and other actors who could not attend. Furthermore, the concepts are presented at the largest international software engineering conference and an open source conference, spawning interest in the research community as well. We conclude from the pre-study that there is a sufficient basis of interested partners to form a continuation project to explore the problems and propose solutions, preferably related to concrete cases of open data collaboration.

1 Introduction

Under the slogan "Data is the new oil"[1], many businesses try to collect as much data they can from their customers, products, and public sources. The slogan, coined in 2006, has become even more valid in current times when computer systems are able to process huge amounts of data, more data than ever is collected through apps and sensors and machine learning and artificial intelligence (ML/AI) are experiencing a renaissance. Data defined components are more and more common in software systems. Access to data can be a competitive advantage, especially for pioneers in the market. However, as time passes and competitors advance, data may turn into a commodity, i.e. an asset that is necessary for the business, but does not bring competitive advantage anymore. Still, data must remain reliable and of high quality to be useful for the companies, which implies costs for maintenance and quality assurance.

Open Source Software (OSS) is utilized in almost any software system, including commercial offerings. Software can be a competitive advantage, but at some point, competition and partners might catch up. Then software becomes a technical debt and a commodity instead of an advantage. Still, the software needs to maintain proper quality. OSS is a means to share platform software and tools with partners and even competitors both to reduce cost as well as to promote open innovation.

Chesbrough coined the term Open innovation, (OI) [2] , initially to refer to exchange of ideas. OI is "a distributed innovation process across organizational boundaries, using pecuniary and non-pecuniary mechanisms" [3] . OSS is one way of fostering OI. Open Data, i.e. public agencies giving access to public data, is brought forward as an enabler for innovation and entrepreneurship, e.g., by Lakomaa and Kallberg [6].

We propose the concept Open Data Collaboration (ODC)[2] as the practices and approaches for sharing data among organizations to promote OI and to share cost with others. Open Data (OD), on the other hand, refers to data made available to anyone. There are both technical and organizational challenges with ODC for data defined software systems; e.g. how to ensure data integrity for individuals when data is shared across organizations; business models and strategies for when to share data and when to keep it as a competitive advantage; technical solutions for sharing data in a secure and efficient way – especially for small devices with limited capacity such as IoT devices. In order to understand the challenges and opportunities facing companies around ODC, we ran a series of focus groups with practitioners. In the focus groups, participants from different companies discussed data, open data and open data collaborations.

The rest of this report is organized as follows: Section 2 presents the project, how we organized the focus groups, and the participants. Section 2.4 presents some more background on OI, OSS, and OD and related work. The main results are found in Section 4. In section 5, we present our suggestions for a research agenda on ODC.

[1] Usually originally attributed to Clive Humby, UK mathematician and architect of Tesco's Clubcard, 2006

[2] The pre-study is called "Open Collaborative Data (OCD)". However, during the pre-study, we decided to call the concept "Open Data Collaboration (ODC)" instead due to acronym interference with other concepts.

2 Project setup

To understand the challenges and opportunities facing companies around ODC, we organized three three-hour workshops, comprising five focus groups. Our goal with the focus group discussions was to get the participants' views and experience of practices around data in general and sharing data specifically.

2.1 Project objective and overview

The project aims at exploring the concept of ODC as an innovation platform for industry to innovate, to share costs and to ensure quality of data for training of machine learning applications. We wanted to understand the challenges companies identify, and which opportunities can help to nudge the companies in the direction of ODC. We see ODC as a complement to OSS and a way to approach OI.

2.2 Project outline and process

Based on the goal to understand challenges and opportunities, we organized focus groups with commercial and public organizations. Open invitations to the workshops were sent out to our professional networks, in which the focus group discussions were held. Participants attended one focus group. In total, we organized five focus groups. Figure 1 outlines the process.

Each of the workshop session followed a similar scheme. First an introduction to the concept of ODC was given. Then the attendants were split into focus groups of 5-7 participants, plus one moderator and one secretary. The focus groups then discussed topics related to ODC under three main themes:

- What type of data does your company/organization use/produce?
- Which data can be shared? Under which conditions? To whom?
- Which are the enablers and hinders for sharing data?

The composition of the focus groups was made based on the affiliation of the participants; company, research organization or public agency, to create a wide variety of participants in the discussion.

During the focus group sessions, we let the participants' scenarios for data drive the discussion as much as possible. Only in cases where we wanted the group to discuss a certain point, we introduced pre-defined scenarios.

After the focus group session, the groups reassembled, and a summary of each group was presented and discussed.

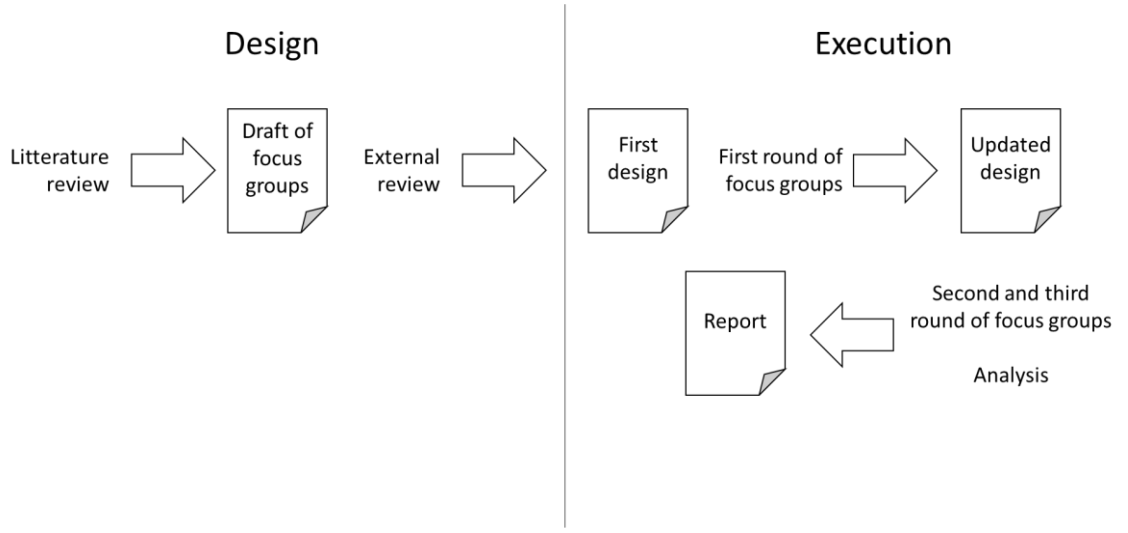


Figure 1 Process overview

After the focus group series, the collected data was analyzed and summarized. The notes were grouped according to main findings. The preliminary findings were presented at a public presentation. We invited both the focus group participants as well as others. There were about 40 persons attending the public presentation.

Based on the focus groups and the feedback from the public presentation of the preliminary results, a final analysis was performed. The main results are reported in section 4. Based on the main findings, we propose several topics for a research agenda, found in section 5.

2.3 Focus group and participants

We ran three workshops with a total of five focus groups in March and April 2019. The preliminary results were presented at a seminar in May 2019.

In total we had 27 participants from 22 companies and public organizations. Table 1 outlines the participants who agreed to have their name and affiliation in the report.

Table 1 Participants

Participant	Organization
Martin Börjesson	2021.AI
Magnus Midholt	ARM
Jiandan Chen	Axis Communications AB
Martin Ljungqvist	Axis Communications AB
Andreas Schön	Cybercom
Mikael Lostedt	Ericsson
Karin Rathsman	European Spallation Source ERIC
Hussan Munir	Lund University

Participant	Organization
Britta Duve Hansen	Lunds kommun
Celine Berggreen-Clausen	Malmö stad
Henrik Svenell	Qlik
Robert Lann	Sensative AB
Lijo George	Sony
Pär Spjuth	Sony
Christer Pihl	T-Kartor
Jonas Söderberg	RISE
Jakub Jurasz	MDH University and AGH University, Cracow, Poland.

2.4 Outreach

The invitations to the project events (focus groups and result presentation) were sent to the professional networks of RISE, Lund university (Digit@LTH), and Mobile heights. The overlap between the networks are significant, but still we estimate the coverage be at least 400 people. As mentioned above, 27 persons attended the focus groups and 40 the final presentation.

In addition, project results are presented at the ICSE New ideas and emerging results track in Montreal, Canada, in May, and at the 15th International Symposium on Open Collaboration, that will take place in Skövde in August. Further, the outcomes were presented at Swedsoft's Industrial Open Source Network meeting in Lund in May.

3 Background and Related work

3.1 Openness

Chesbrough coined the term Open innovation (OI) [2] , initially to refer to exchange of ideas. OI is "a paradigm that assumes that firms can and should use external ideas as well as internal ideas … as they look to advance their technology." Later, Chesbrough et al. redefined OI as "a distributed innovation process across organizational boundaries, using pecuniary and non-pecuniary mechanisms" [3] . In their systematic literature review on OI in software engineering, Munir et al. identified nine research themes, including OI strategies, challenges, benefits, communities, management, and intellectual property (IP) strategies [8] . However, none of the topics relate to open data, in the sense of sharing data across organizational boundaries.

OSS in commercial business has emerged to share platform software and tools with collaborators and competitors. While OSS in the 1980's was more a philosophical and

political issue, it turned in the 1990's into a commercial phenomenon, through Linux and free BSD3. Studies on open software tools [9] as well as on product software [7] indicate that OSS plays a key role for software business, although it has to be managed accordingly.

Open Data is brought forward as an enabler for innovation and entrepreneurship, e.g., by Lakomaa and Kallberg [6] . However, this refers to public agencies opening up their data to private companies, not – like in the OSS case – companies sharing between them. Susha et al. developed a taxonomy to describe the variation in such Open Data [12] . Lakomaa and Kallberg indicate that there is a cost for the agencies to release Open Data, and that it is a political decision to take that cost to catalyze innovation [9]. We have not found any research on companies sharing data between organizations.

3.2 Sharing data

Data is a key asset in systems-of-systems (SoS) [1] . In a SoS, several organizations cooperate to deliver a value which a single organization cannot deliver by themselves. Exchange of data is inevitable. We see several challenges, both technical and legal, for how to handle data in this context, where the configuration of the system is not always known at development time and can vary frequently during operation. In project such as PIMM DMA[3] – a systems-of-systems around a closed-pit mine – data has emerged as a major challenge to achieve an efficient innovation collaboration. There is a combination of business concerns (should we share), legal concerns (can we share), and technical concerns (how do we share).

There are initiatives on sharing data for machine learning, like Open Mined[4]. They apply the principle of Federated learning, meaning that the ML model is brought to the data, rather than bringing the data to the model. This helps addressing the privacy issues partly, but introduces a transparency problem, in that you don't know which data the ML model is trained on.

Frizzo-Barker et al [5] conducted a systematic mapping study of research on Big Data in business scholarship. With respect to openness, they only identify open collection of data (crowdsourcing) and open tools for big data analysis. Among the challenges identified in the primary research, the top challenges are i) how to take advantage of the enormous volumes of data, ii) handle the risk of privacy and ethical infringements, and iii) managing the cost-benefit trade-off "of using big data for decision-making, the validation and integrity of collected data, and the complexities of dealing with highly distributed data sources." Hence, the costs are identified, but no solutions discussed.

Del Vecchio et al [4] provided an extensive overview and analysis of research on the borderline between information systems and innovation management, with focus on open innovation. They report how open source platforms, such as Hadoop, contributes to open innovation based on Big Data, and how analysis of data may lead to business innovation. They also touch upon using open data, scraping the web etc. However, they do not find any research on the possibility to share data between corporations to foster

[3] www.sics.se/projects/pimm

[4] https://www.openmined.org

open innovation. This is the gap our proposal on Open Data Collaboration (ODC) aims to fill.

3.3 Related initiatives in Sweden

There are several projects and initiatives related to ODC. The list below contains some of those.

- AI Innovation of Sweden is running a data factory to accelerate research and innovation by making data more available https://www.ai.se/en/resources/data-factory
- Swedish Agency for Economic and Regional Growth (Tillväxtverket) is running a platform for open data and data-driven innovation. http://challengesgov.se
- The agency for Digital Government (Digitaliseringsmyndigheten DIGG) has the mission to facilitate open data from the public sector. https://oppnadata.se
- The Vinnova funded project Sjyst data aims at facilitating trust and business development in an ethical manner. https://sjystdata.se//
- The Stockholm region and municipalities are running a project to improve the infrastructure for open data, http://smartsthlm.stockholm.se/2018/01/09/15-miljoner-kronor-till-kraftsamling-for-oppna-data-i-stockholmsregionen/

We have informed these agencies about the existences of the pre-study and invited them to the focus group meetings and the final presentation, although they did not attend.

4 Results from the focus groups

Based on notes from the focus group discussions, we identified seven themes related to data, which we explore below, related to:

1. Open data strategies
2. Data purchase
3. Mindset for sharing
4. Costs for data
5. Quality and trust
6. Standards and APIs
7. Competence

The themes cover a broad range of topics, technical, to organizational, legal, and cultural. Furthermore, these aspects are intertwined, indicating that they have to be addressed in cross topic fashion, rather than one by one.

Open data strategies are uncommon.

> The concept of ODC and strategies for ODC are still in their infancy. Existing literature only addresses open data, as shared by public organization, and thus does not give support in defining strategies and processes for ODC, which primarily refers to private organizations sharing data.

All participants voluntarily participated in the workshops, indicating an interest in data as part of their business. Many of the participants collected data and some also have data intensive components – such as machine learning – as part of their products or systems. However, none of the participants actively worked with sharing data nor had any defined process or strategies for data in relation to open innovation.

As a conclusion from the focus groups, the business value for the organizations is a precondition for sharing data. This far, no company had explicitly defined an open data strategy, while several companies had a corresponding strategy for open source software.

Data is not always possible to purchase.

Certain type of data can be purchased, such as market data and data collected e.g. by smart phones and apps like Facebook. There exist marketplaces and data brokers for this. Furthermore, there are open initiatives, e.g. openmined[5], to share data.

However, even if a company wants to purchase data, e.g. annotated image data for machine learning, there are a lack of available resources. Some type of data is not and will likely not be available to buy. Often, companies are required to team up with others, perhaps even competitors, who are also collecting similar data to get access to more data.

In the second workshop, participants speculated that there need to be public initiatives to build large data sets. The platforms, such as Google and Facebook, have lots of data but they control it. Furthermore, for others to catch up on technology leaders in a certain domain, e.g. to Tesla, companies need to cooperate as the large platform companies have a head-start. This was suggested by participants in the second workshop.

Participants in the third workshop believed it is difficult for data brokerage to have a sustainable business, which explains the lack of options.

Sharing data requires a mindset change.

Open Innovation, whether through OSS, ODC or other mechanism, entails opening up key processes to others and potentially giving away assets. The idea is that the long-term competitiveness is improved even though short-term it can seem as if a competitive advantage is lost. The change of mindset is not always easy, which focus group participants illustrated by referring to their process of turning open source.

Participants in the second workshop specifically mentioned that mutual sharing is key. That is, to be an ODC partner, you must give something away to receive something back. In the third workshop, participants also pointed out that there needs to be a business rationale internally to motivate investments in sharing.

[5] https://www.openmined.org

There are costs of working with and for sharing data to consider.

Collecting useful data and ensuring its quality for the purpose intended requires investments. Data often need to be processed – not seldom by humans – to be useful. There might be additional costs related to sharing of data, e.g., to ensure reliable and secure communication, to add mechanisms to filter out which data to share, etc.

Annotation is mentioned by the participants in the first and second workshop as a costly and labor-intense process. If the data is also being shared as open data, additional resources are needed to validate and distribute the data. Participants in the second workshop mention that their systems are not prepared for sharing – neither APIs nor content.

Quality and trust are key issue.

As data becomes more and more important for successful development and reliable operation, the requirements on quality of the data increases. Similar to ensuring the quality of the software, data also needs to be quality assured. Furthermore, just as reliable communication can be key for a system to operate as intended, data also needs to be reliable.

Participants in the second workshop mentioned that having multiple sources of data can improve quality as well as sharing of costs related to the data. Furthermore, if more companies are using the same data, inaccuracies are more likely to be discovered. Depending on the type of data, sometimes quality is about providing an exact fact – e.g. a certain label – while in other types of data, particularly measurement data, averaging over several sources gives more robust input.

The participants further pointed out that there needs to be a trust among the sharing parties. If an external party is responsible for the quality assurance and the relationship is non-pecuniary, trust needs to be established by other means. Lastly, trust needs to be fostered and maintained.

There is a need for standards and well-defined APIs.

Sharing of data puts requirements on the technical infrastructure. If data should be shared with several different organizations and over time, there is a need for standardization of formats, APIs, etc. for it to be technically realistic and cost-efficient.

The participants mention that many of their systems are not prepared for sharing data. Furthermore, they also mention that even if sharing is technically possible, it is also required that the procedures for collecting and processing the data are standardized to ensure data is interpreted the same by different organizations.

Working data driven and with open data requires new competences.

Analyzing data and making it an integral part of the business requires new competences such as data analysts as well as a general understanding for

several roles of how to use data. Sharing data adds additional needs for understanding licenses around data and integrity issues, specifically in the light of GDPR.

Participants in the second workshop pointed out that both the operational layer (those doing the actual work) and the strategic layer (those with power to decide) need to be aligned and understand data and sharing. In the third workshop, liability was brought up both in terms of liability if sharing data with faults or issues in it and liability when using data with faults or issues in it.

5 A research agenda

Most companies are aware of the usefulness of data and several have already come a long way in using it in their operations. Some have also realized that it is not easy to collect, analyze and curate the data. However, even though we got many participants in the focus groups, the concept of open data collaborations was largely new to the participants.

We postulate that the cost of curating and maintaining data will sooner or later exceed its business value [10] . However, there will be large differences in different domains and ecosystems around the world. Based on literature and our findings from the focus groups we propose a research agenda organized in technical, organizational, business and legal aspects.

5.1 Technical

Sharing of data requires an infrastructure and mechanisms. There are needs within the organizations to have a platform where data can be managed to allow for the appropriate access control, communication enablers such as APIs and standardized formats, as well as versioning of data to ensure traceability. If sharing occurs on a larger scale, a proper technical infrastructure is key to keep administrative cost down.

Today there is a lack of suitable solutions. We see a need for open platforms where both source code as well as the hosting of the platform can itself be open and shared among many organizations. Open standards are another interesting area to further understand.

5.2 Organizational

For an organization to share data in a professional and high-quality manner, there needs to be processes and competencies in place. To assess the readiness for ODC in organizations, a maturity model could help guiding the development. Research topics include to understand the appropriate steps to increase the maturity level and how that aligns with ODC concepts. There might also be a need for new roles and new competences in the organizations.

5.3 Business

Any organization need to assess the value of their activities. As ODC is a new concept, there is a lack of models to estimate value as well as costs. This is also related to strategies

for ODC. Sometimes it might be most beneficial to be a leader and invest a lot and share a lot, while sometimes it is better to be a follower and try to keep costs down. Similar to OSS and OI, the criteria and trade-offs needs to be better understood to support business decisions. It is also important to have a lifecycle perspective as any sharing component that is included might require an investment for a long time.

Some of these aspects may be based on practices for OSS and OI. There is an overlap with software ecosystems and other community practices around OI. However, there are also unique aspects with ODC which need further research to be better understood.

5.4 Legal

There are legal aspects around data whether sharing or receiving data. There exist license models for data but as we have seen with OSS, this is a complicated matter. Furthermore, privacy and integrity of persons and organizations needs consideration. Compliance to laws like GDPR needs to be considered from the start to avoid issues.

Liability might also be impacted. If you are sharing data, depending on the license and the agreement of the ones using your data, liability might remain with you. Important to note is if your agreements even indicate that you need to continue to provide data and how that affects your business decisions. If you are receiving data, you also need to be aware of liability, e.g. what happens if you no longer get data?

6 References

[1] J. Axelsson. Systems-of-systems for border-crossing innovation in the digitized society-A strategic research and innovation agenda for Sweden. (2015), http://www.diva-ortal.org/smash/get/diva2:1043518/FULLTEXT01.pdf

[2] H. W. Chesbrough, Open innovation: the new imperative for creating and profiting from technology. Boston, Mass.: Harvard Business School Press, 2003.

[3] H. Chesbrough, W. Vanhaverbeke, and J. West, Eds., New Frontiers in Open Innovation. Oxford University Press, 2014.

[4] P. Del Vecchio, A. Di Minin, A. M. Petruzzelli, U. Panniello, and S. Pirri. Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. Creativity and Innovation Management, 27(1):6-22, 2017.

[5] J. Frizzo-Barker, P. A. Chow-White, M. Mozafari, and D. Ha. An empirical study of the rise of big data in business scholarship. International Journal of Information Management, 36(3):403-413, 2016.

[6] E. Lakomaa and J. Kallberg, "Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs," IEEE Access, vol. 1, pp. 558–563, 2013.

[7] J. Linåker, H. Munir, K. Wnuk, and C. Mols, "Motivating the contributions: An open innovation perspective on what to share as open source software," Journal of Systems and Software, vol. 135, pp. 17 – 36, 2018.

[8] H. Munir, K. Wnuk, and P. Runeson, "Open innovation in software engineering: A systematic mapping study," Empirical Software Engineering, vol. 21, no. 2, pp. 684–723, 2016.

[9] H. Munir, P. Runeson, and K. Wnuk, "A theory of openness for software engineering tools in software organizations," Inf. and Softw. Technology, vol. 97, pp. 26–45, 2018.

[10] P. Runeson, Open Collaborative Data – using OSS principles to share data in SW engineering. In ICSE (NIER), pages 25–28. IEEE / ACM, Toronto, Canada, May 2019.

[11] R. H. Coase. 1937. The nature of the firm. Economica, 4(16), pp. 386–405, 1937.

[12] I. Susha, M. Janssen, and S. Verhulst, "Data collaboratives as a new frontier of cross-sector partnerships in the age of open data: Taxonomy development," in 50th Hawaii Int. Conf. on System Sciences, HICSS. AIS Electronic Library, 2017.

RISE Research Institutes of Sweden AB
Box 857, SE-501 15 BORÅS, Sweden
Telephone: +46 10 516 50 00
E-mail: info@ri.se, Internet: www.ri.se

SICS
RISE Report 2019:77
ISBN: 978-91-89049-06-2