Challenges and Opportunities in Open Data Collaboration – a focus group study

Runeson, Per; Olsson, Thomas

Published in:
46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

DOI:
10.1109/SEAA51224.2020.00044

2020

Document Version:
Peer reviewed version (aka post-print)

Challenges and Opportunities in Open Data Collaboration – a focus group study

Accepted for publication in Proceedings 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)

Per Runeson and Thomas Olsson

Abstract

Data-driven software is becoming prevalent, especially with the advent of machine learning and artificial intelligence. With data-driven systems come both challenges – to keep collecting and maintaining high quality data – and opportunities – open innovation by sharing data with others. We propose Open Data Collaboration (ODC) to describe pecuniary and non-pecuniary sharing of open data, similar to Open Source Software (OSS) and in contrast to Open Government Data (OGD), where public authorities share data. To understand challenges and opportunities with ODC, we organized five focus groups with in total 27 practitioners from 22 companies, public organizations, and research institutes. In the discussions, we observed a general interest in the subject, both from private companies and public authorities. We also noticed similarities in attitudes to open innovation practices, i.e. initial resistance which gradually turned into interest. While several of the participants were experienced in open source software, no had shared data openly. Based on the findings, we identify challenges which we set out to continue addressing in future research.

1 Introduction

Open innovation and co-creation is a way for organizations to leverage the creativity outside the own organization [1]. Open innovation is not new in software engineering; open source software [2] and software ecosystems [3] are examples of how open innovation can be fostered. Data in general and big data in particular has for the last decade become prevalent [4, 5], in particular as data-driven systems and machine learning (ML) applications require lots of high-quality data. Raj et al. point to data management challenges such as shortage of data, need for sharing techniques, and data quality [6]. As suggested in our previous work, there is a need to adopt co-creation and collaboration principles to harness the innovation potential in the age of data and to manage costs [7]. This is in line with other researchers who see the need for an ecosystem strategy when working with open data [8].

Open Source Software (OSS) is utilized in almost all software systems, and is integrated with commercial offerings. OSS is a means to share platform software and tools with partners – and even competitors – both to reduce cost and promote open innovation. Chesbrough coined the term Open Innovation (OI) as "a distributed innovation process across organizational boundaries, using pecuniary and non-pecuniary mechanisms" [9]. Open Government Data (OGD), i.e. public agencies giving access to public data, is studied quite intensively [10] and brought forward as an enabler for innovation and entrepreneurship [11, 12]. Just recently, the Bennett Institute for Policy, Cambridge, launched a report on "The Value of Data" [5] with a focus on public policy for data. However, as far as we have seen, the opening of data between commercial organizations to create more value, with or without governmental involvement, is not practiced to any major extent.

To advance this field, we wanted to explore practitioners' views on collaboration about data in an open innovation setting. We therefore launched a focus group series on attitudes and practices around collaborating with external organizations on data. We have previously proposed the term "Open Data Collaboration" (ODC) [13] when discussing preliminary results of the focus group study. There are both technical and organizational challenges with ODC. For example, adhering to privacy laws when data is shared across organizations, business models and strategies for when to share data and when to keep it as a competitive advantage, and technical solutions for sharing data in secure and efficient ways – especially for small devices with limited capacity such as IoT devices. Our preliminary results are presented in a technical report and a poster presentation [13, 14]. In this paper, we report the full analysis of the focus group study and outline consequential further work.

The rest of this paper is organized as follows: Related work is presented in Section 2. The research

method is outlined in Section 3. Section 4 presents the results and Section 5 the discussion and implications of the findings. Threats to the validity are presented in Section 6 and the paper is concluded in Section 7.

2 Related work

Raj et al. point out that the quality of the outcome of machine learning in general and deep learning in particular, is related to the data used to train the systems [6]. In their multiple-case study, they identify, for example, shortage of data (not being able to train a system fully), need for sharing and tracking techniques (using unclear input to train, makes it difficult to reproduce results), and data quality (a trained system is only as good as the data it trained on). Our work complements this work on data management challenges for machine learning with a perspective on data sharing practices.

Collaborating on OSS is a well established practice with supporting theoretical knowledge. Alves et al. surveyed research on governance of software ecosystems and found 89 relevant papers [15]. They observe the importance of the platform owner and balancing of rights between owners and contributors. Our own research on Open Tools ecosystems [16] and product features in software ecosystems [2] focus on strategic choices on contributions, as a means for influence.

Attard et al. [10] systematically surveyed literature on OGD and synthesized 75 papers with focus on governments as actors. The primary goal of government authorities is to increase transparency, although the access of information as such is brought forward as a benefit. However, the involvement by private companies or citizens as data providers is not addressed. Attard et al. identify five categories of challenges for OGD, 1) technical, 2) policy/legal, 3) economic/financial, 4) organizational, and 5) cultural, which seem to be relevant also for ODC.

The potential innovation benefit from OGD ecosystems is discussed by Zuiderwijk et al. [17]. They advice how to create OGD ecosystems and define four key elements of an OGD ecosystem: 1) government data provisioning, 2) data access and licensing, 3) data processing, and 4) feedback to data providers. Further, to get ecosystems into function, three additional elements are defined: 5) usage examples, 6) quality management system, and 7) metadata. A survey among entrepreneurs indicate significant interest in OGD [11]. However, the sustainability of funding is a threat to such entrepreneurial initiatives [18]. Case studies of OGD, e.g. by Dawes et al. [12] and Styrin et al. [19], indicate varying practices in different countries, and stress the socio-technical character of OGD ecosystems.

The Bennett Institute for Public Policy recently published a report on "The Value of Data" [5]. It is also primarily focused on OGD, but broadens the view by discussing trade-offs when sharing data. *"Value comes from data being brought together, and that requires organisations to let others use the data they hold. But if they do, organisations will not get all the benefits from data they have collected, and perhaps not enough benefits to cover the cost of collecting and storing the data in the first place."* They discuss market and non-market solutions to estimate value on data, and point to the need for intermediary organizations to handle data exchange, e.g. trusts. Further in their Data Spectrum model, they indicate that data may gradually be transferred in stages: closed–shared–open.

In the literature, we did not find any research on the possibility to share data between corporations as a means to foster open innovation. That is where our proposal on Open Data Collaboration (ODC) aims to fill a gap.

3 Research Method

To study the phenomenon of ODC, we conducted an exploratory qualitative survey study [20]. As the concepts are new and we are interested in their relevance to practice, our first step is to explore practitioners' views. We wanted to understand different organizations' attitudes to challenges and opportunities with ODC. Thus, we strived to get a broad range of organizations represented in our study. Further, as the study is exploratory, we rather wanted to survey many sources superficially, rather than going into depth in fewer cases.

As a consequence, we choose focus groups as our method for data collection [21]. We invited participants broadly from our extensive network of commercial and public organizations to attend workshops in three different locations – Lund, Gothenburg, and Stockholm, two time slots in each place.

The overarching research questions for our study are derived, based on earlier research on OSS and our hypotheses on potentials for ODC [7]:

RQ1 What data is produced and used within and shared among the organizations?

RQ2 What are the attitudes towards sharing data in an ODC fashion?

RQ3 Which are the expected challenges and opportunities with ODC?

3.1 Participants

To understand the attitudes, challenges, and opportunities facing organizations on ODC, we invited participants to focus group meetings through three main networks; the regional ICT innovation cluster organization[1], the university's collaboration network, and the research institute's corresponding contacts. Attendants could register for any of six workshop occasions in three different locations.

Based on the registration pattern, we organized three workshops in two locations with 27 participants from 22 companies, public organizations, and non-profit organizations (see Table 1). In two of the workshops, we split the attendants into two focus groups, thus running in total five focus groups. The focus group meetings were organized within three weeks time in March and early April 2019. Below, we refer to them as FG1.1, FG1.2, FG2.1, FG2.2, and FG3.

The 15 private organizations serve in different domains, as presented in Table 2 and there was a mix of large, medium, and small enterprises. Even though participants are sampled by convenience, and thus not a representative sample in statistical meaning, they represent a broad range of industry and public organizations.

The participants had different roles in their organizations. In general, attendants were senior people and had technical or middle management roles. Most organizations had only one representative, although some large enterprises sent more than one attendant.

In the workshops with two parallel focus groups, we split them into groups of size 5-8 by area of interest, for example, health or automotive. Further, we tried to avoid having several persons from one organization in the same focus group. In addition, each focus group had one moderator – one of the researchers – and one secretary – in four out of the five focus groups an external person, while in one focus group (FG3) the first author was the secretary.

3.2 Data collection

We collected data through the workshops and validated the initial findings in a public event. Each of the workshop sessions followed a similar scheme. We first broadly introduced the concept of ODC. We then split the attendants into focus groups. These then discussed topics related to ODC under three main themes:

- What type of data does your organization use or produce?
- Which data can be shared? Under which conditions? With whom?
- Which are the challenges and opportunities for sharing data?

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 our own example scenarios. In conjunction with the focus group sessions, the two groups reassembled, and a summary of each group was presented and discussed. The schedule for the focus group can be found in Appendix 7.

We presented our preliminary findings at a public event with 40 attendants, where both the participants in the focus groups and others were invited. We planned the event to be interactive where the participants were asked to confirm our interpretations. The participants were also encouraged to ask questions and make comments. We also organized an informal mingle before and after the presentation, to elicit additional reflections and comments.

3.3 Analysis

Meeting notes were taken by the secretary in the focus group meetings. The findings from each focus group were briefly summarized after each workshop, structured according to the questions of the focus group.

After all the workshops, all notes were merged and then coded [20]. The coding and grouping of codes was performed by one of the researchers. We started the analysis with *a priori* codes, based on the topics for the focus group meeting (see Appendix), namely:

C1) Organizational characteristics

C2) Data characteristics

C3) Data sharing conditions

C4) Challenges

C5) Opportunities

Next, the coding was refined and synthesized in two iterations, into eight final topic codes, see Table 3.

After this process, conducted by the second author, the code structure was reviewed by the first author. Changes in the outcome were primarily about modification of terms and a more precise definition of the codes. The final codes and their definitions are presented in Table 3. The results are presented and structured according to the final codes in the next section.

[1] `http://mobileheights.org`

Table 1: Overview of participation from different organizations

	Type of organization	Number of participants
Public	University or research organization	4
	Municipality	3
Private	Non-profit organization	1
	Small or medium enterprise	8
	Large enterprise	11 (from 6 org's)
	Total	27 (22 org's)

Table 2: Overview of domain of the private organizations

Domain	Number of org's
Automotive	1
Computer and chips	1
IoT	3
IT consultant	3
IT services	3
Medtech	1
Telecom	3
Total	15

4 Results

The following sections first present types of data used and produced within and between the organizations (RQ1). Then we present the main findings on attitudes to ODC (RQ2), structured according to the final codes, as defined by F1–F8 in Table 3. Findings in terms of challenges and opportunities (RQ3) are also discussed for each code.

4.1 Types of data

In the focus groups several categories of data were identified, both based on application domain and data characteristics. We identified seven broad categories: Maps, Society, Position, Images, Sensors, Human, and Business.

Map data can be general, physical maps, with different layers of information. OpenStreetMap[2] is an example of an ODC for maps. There are also companies, making business on map data (e.g. for military purposes) and there is a Swedish governmental authority that also partially is business based[3].

Society data includes all kinds of data related to the society. It may partly be seen as an extension to map data about buildings or technical infrastructure. Society data may also be dynamic, related to heat, electricity and different aspects of communication. However, it may also include information about events, regulations, decisions, as well as statistics on population or economical aspects of the society.

Position data is related to maps, but is focused on the dynamics of transportation and individual movements. These can be seen as snapshots at a certain point in time, or time series for historical analyses and prediction.

Images are data for training of machine learning applications. Faces and plants, were mentioned as examples for different machine learning applications.

Sensor data refer to different kinds of measurement data from sensors, such as temperature, light, humidity in the environment, or sensors in a control system of a production plant. Sensors may be fixed or moving, the latter e.g. in a vehicle, where it may be connected to position data.

Human data is the most sensitive type, as it may be connected to many other types of data. Particular examples include behaviour, position, and health data.

Business data may similarly be sensitive, as it includes customer data or is about the business as such. This data may also include usage data on the product/service provided.

In addition, the focus groups (especially FG2.1) discussed synthetic or generated data, as a source of data of different types for training of machine learning applications. An overview of which focus groups represented different types of data can be found in Table 4.

4.2 Business of data

The focus group participants expressed a sober attitude to the value of data, in contrast with evangelist statements on "data as the new oil". Participants clearly stressed that data have no value in itself, but must be connected to some business. One participant expressed that *"big data means nothing if you do not have a business value"* [FG2.2]. An-

[2] `www.openstreetmap.org`

[3] `https://www.lantmateriet.se/en/`

Table 3: Final codes

ID	Code name	Code definition
F1	Business of data	Potential business models, costs related to the collection and annotation of data, and business value of data
F2	Business of collaboration	Conditions for and effects of collaboration around data
F3	Data acquisition	Acquisition and brokerage of data
F4	Relationships	Relationships between parties sharing data
F5	Competition	Aspect of competition between parties sharing data
F6	Quality	Quality of data, and what contributes to the quality
F7	Maturity	How a data ecosystems may mature, with a particular focus on competence needs and standardization
F8	Legal	Licensing and legislation for data

Table 4: Data types as discussed by the focus groups, respectively

Data type	FG1.1	FG1.2	FG2.1	FG2.2	FG3
Maps	x				
Society	x		x	x	
Position	x	x	x		x
Images	x	x	x		
Sensors	x	x	x	x	x
Human	x	x	x	x	
Business	x		x		

other one expressed that it was difficult to put a number on the value of data [FG1.2]. This conservative position is particularly interesting as we invited participants to the workshop with focus on data – i.e. a clear bias for an interest in data.

Several participants expressed that usage data is important to improve their products and services [FG1.1]. All organizations in the workshops do collect data in one way or another.

The opinions on 'spillover' data differed, i.e. data which is not intentionally collected but gained as a by product of other data collection. Some argued for this data being well suited for sharing or selling, while another participant noted that the 'gems' can be found in the part of the data that you did not intentionally collect [FG2.2]. It was noticed that *"Google has a broad business model so they can cross-fertilize domains"* [FG3], taken as an indication that their success is a kind of internal harvesting of spillover data.

Annotation of data is mentioned by the participants as a costly and labor-intense process [FG1.2]. Having access to annotated data is key for machine learning. The participants see an opportunity to collaborate with other organizations in the annotation efforts.

4.3 Business of collaboration around data

There are costs related to collecting data and ensuring its quality for the intended purpose. Data often need to be processed – not seldom by humans – to be useful. There might be additional costs related to data sharing, e.g., to ensure reliable and secure communication as well as additional mechanisms to filter out which data to actually share. Participants mentioned that *"their systems are not prepared for sharing data – neither with respect to APIs nor to content"* [FG2.2]. Furthermore, if the data is being shared as open data, additional resources are needed to validate and distribute the data. Hence, the participants agreed that collaborating around data entails costs which needs to be matched by getting something in return. Collaboration without business value will not happen [FG3].

Sharing data within an organization may also be a challenge. One of the municipalities participating in the workshop, devoted it to political factors rather than technical ones [FG1.1]. That is, structures, regulations, and ways of working become challenges for sharing data even within a municipality. They are, however, sharing "master data", i.e. information about inhabitants, addresses, and similarly. Contrasting this, there are certain legislation requiring municipalities to share data with Lantmäteriet[4] and at the same time are required to pay for data from them as well [FG1.1]. In this case, the legislation is a challenge for ODC.

Participants pointed out that collaborating around data might improve the quality of the data [FG1.2]. For example, if data shared with others is being annotated, this might add value. One participant pointed out, however, that not all data is equally interesting to collaborate around. They hy-

[4]A governmental authority that maps the country, demarcates boundaries and helps guarantee secure ownership of Swedens real property.

pothesize that more general data is of greater value for collaboration, as opposed to very specific data [FG2.2]. Another participant mentioned that collaborating around data can be a way to increase market presence as well – by getting insights and thereby the ability to build products and services for new customers [FG1.2]. Another potential opportunity is the pillar of open innovation – by giving away some asset, the total market of the collaboration partners becomes bigger, a "win-win situation". This, however, requires adoption of open innovation principles.

One participant stated that they might be more inclined to *"trading the data with someone who is not a competitor"* [FG1.2]. Many participants reported having a concern that they give away a business value when collaborating on data. Therefore, they would rather collaborate with organizations that are not direct competitors.

4.4 Data acquisition

Certain types of data can be purchased, such as market data and data collected by smart phones and apps. Marketplaces and data brokers exist, even though participants had seen examples *"especially in the insurance business and it was hard to get them fly"* [FG3]. However, even if a company wants to acquire data, e.g. annotated image data for machine learning, there is 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 [FG2.2]. The platform companies, 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, companies need to cooperate as the large platform companies have a head-start.

4.5 Relationships

The participants pointed out that there has to be a trustful relationship among the collaborating 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.

Participants in FG2.2 specifically mentioned that mutual sharing is key. That is, to be an ODC partner, you must give something away to receive something back. In FG3, participants also pointed out that there has to be a business rationale internally to motivate investments in sharing.

Collaborating around data implies that data might be owned by other organizations. A participants stated that data that *"owning your data you know it is correct"* [FG2.1] and thus you may be sure it is more reliable. This implies that the more important the data is for your business the higher is the risk if the data is not owned by your own organization.

4.6 Competition

Competition and competitors is a theme that recurred several times in the focus groups. One participant suggested that a way for smaller organizations to compete on a global market is to collaborate on data [FG2.2]. Otherwise, the large multinational companies will have a too large advantage as they can collect and curate much more data. They suggest that forming local and regional clusters of collaborators may give an advantage.

Another participant suggested that making data publicly available is another way of taking away the competitive advantage and, at the same time, contribute to the overall greater good for the society [FG2.2].

One hindrance for collaborating – a participant in the FG3 mentioned – might be if the other organizations are better at turning the data into business value. Hence, it might be a disadvantage to collaborate on data or making it publicly available if other organizations are perceived as being faster.

4.7 Quality

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

Participants mentioned that having multiple sources of data can improve quality as well as sharing of costs related to the data [FG2.2]. 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.

Transparency also came up as a topic in the second workshop [FG2.2]. To be able to trust data, it must be transparent how data were collected and curated, including any algorithms used in the processing. As opposed to OSS, it is not reasonable that the data is reviewed in its entirety. Rather, the procedures around the data should be checked.

Even though we are used to fast internet connection and cheap storage, it was brought up that the amount of data is growing fast [FG2.1]. Hence, there might be several practical challenges to sharing data as the amount of data grows. For some data, it might also be essential to have up to date data, which further aggravates this challenge.

4.8 Maturity

Sharing software as OSS is an established practice, while ODC is in its infancy. In addition to the cultural resistance to open innovation as part of ODC, participants mention that many of their systems are not prepared for sharing data [FG2.2]. Furthermore, they also state 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.

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 [FG2.2].

The lack of maturity was also brought up, in that several participants were missing suitable standards or APIs for data sharing [FG1.1 and FG2.1]. The municipalities, for example, mentioned that data is stored differently in different municipalities and the technical platforms and APIs also differ. Other organizations had similar observations. Furthermore, organizations are not used to sharing data with others. Hence, there are no processes or procedures for how to act in a collaborative setup [FG1.1], which is an organizational challenge.

4.9 Legal

Legal aspects discussed in the focus groups were primarily related to GDPR[5] and uncertainties about how this regulation will be implemented [FG2.1 and FG2.2]. The uncertainty leads to a challenge, as collaborations might not happen when there is a reluctance to risk legal complications.

[5]The General Data Protection Regulation (EU) 2016/679 is a regulation in EU law on data protection and privacy, which strengthens the right of the individual to its data. http://data.europa.eu/eli/reg/2016/679/oj

There are also issues on license models for data. We have seen with OSS that this is a complicated matter. Liability might also be impacted. If organization share data, depending on the license and the user agreement, liability might remain with the original data providing organization. Further, some participants perceive that legal uncertainties are more of a challenge than technical ones [FG2.2]. Especially public organizations expressed more hesitance due to legal woes [FG2.2].

5 Discussion

The concept of and strategies for ODC are still in their infancy. Existing literature mostly address open data, as shared by public organizations – Open Governmental Data (ODG) [10,17] – and thus does not give support in defining strategies and processes for Open Data Collaboration (ODC), which addresses a wider range of issues, such as business relationships and legal matters.

RQ1 What data is produced and used within and shared among the organizations?

Some data is of rather static nature – e.g., map data – whereas other data is changing all the time – e.g., traffic conditions. This has a major impact on all aspects of sharing. Sharing static data can be manual – physically sending a hard-drive – whereas changing data requires a communication infrastructure – including quality management, monitoring, etc. Especially with large volumes of data, this becomes an important determinant.

The business perspective is also very different for one–off vs. continuous data sharing. Receiving batch data is a one-time investment whereas continuously sending and receiving data – presumably for proper operations – requires a very different business consideration. For example, what is the cost in years from now or what happens if the partner stops providing data?

Privacy is a key concern when discussing data – although peoples' practice still seem to be very relaxed towards sharing data through commercial platforms. However, seen from the perspective of a private company or a government authority, privacy issues seem to be taken very seriously.

Lack of standards and technical infrastructure were often mentioned as a reason why data is not shared. However, data will always be diverse as well as technologies. Furthermore, data change over time, even if the underlying software might not. There are some data marketplaces and broker platforms, although it seems as if they are either geared

towards selling off-the-shelf data – such as marketing data collected from large analytics platforms – or towards open governmental data – where data is publicly available. It seems to us that there is a lack of suitable solutions for organizations to collaborate around data.

RQ2 What are the attitudes towards sharing data in an ODC fashion?

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

Even though the participants are interested in ODC, they were clear that there has to be a business incentive to collaborate. Yet none of the organizations have established business relationships with other organizations around data. This might imply that the organizations – primarily the private ones – have difficulties analyzing and explaining the business value.

The concern of giving away a business advantage was mentioned several times. For example, other organizations might be faster at developing their products and services or that other organizations might find business value in the data which you did not find. Therefore, they are more inclined to work with organizations which are not competitors. This concern is similar for any open innovation and something that was – and sometimes still is – common for open source software. While the core of OCD is open innovation, we believe that practices for collaborating around data is different than for OSS, and hence, need to be better understood to give organizations systematic approaches to evaluate this challenge.

RQ3 Which are the expected challenges and opportunities with ODC?

A main conclusion from the focus groups, is that the business value for the organizations is a precondition for sharing data. None of the participating organizations had explicitly defined an open data strategy, while several companies had a corresponding strategy for OSS. The interest was still great at the workshops, indicating a potential to the organizations. We believe there is an opportunity if public funding can be used as seed money, to allow private and public organizations to engage in ODC, without requiring a direct return of all the investments.

Reduction of costs are also an opportunity to start with ODC. Primarily, the costs related to improving data quality seem to be most important rather than the cost of getting more data. For example, annotation is costly but also data processing to remove noise, etc. We believe that as many organizations now are starting to invest a lot in machine learning and thereby in data, the motivation to engage in data collaborations will increase as the costs will increase. We think this will be particularly valid when data and data-driven software gradually becomes commodity [7].

Several organizations have experience from OSS. One issue that has been challenging for many is licensing, where lack of understanding leads to restricted use. For data, GDPR is also a factor which adds to the legal aspects of ODC. We believe this might, in part, be related to the focus group participants' background. The participants were mainly engineers and technical people. However, the topic of what is allowed and what the legal consequences might be of incidents, was raised several times. Hence, we see a clear challenge that ODC is a concept not yet familiar to the business nor the legal side of organizations.

Furthermore, liability is unclear. What can be expected in terms of not only complying with license and privacy laws but also consequences if, e.g., incorrect data leads to problems when shared with others? We believe here are long-term policy questions to be addressed, as well as needs for creating an environment where it is accepted to take a legal risk and venture into uncharted territory.

Large companies and countries can invest heavily in machine learning – which still has a lot of potential. However, smaller companies and public organizations in smaller countries might not be able to neither invest in the competence needed nor acquire needed data for ML. Hence, there is an opportunity if ODC can be established and open innovation fostered. We hypothesize an even greater innovation potential when organizations are forced to cooperate and share assets. This, we believe, will lead to more open solutions with the citizens in focus rather than lock-in and protectionist approaches, often seen in large tech organizations.

Trusting data sources, and trusting other organization to use shared data in a proper way is a concern for many of the participants. Building trust in a network is hard, and thus some hypothesize that there is a need for a centralized function to ensure the quality and reliability of data. For open government data, the government authority is the guarantee, while in peer to peer sharing we have

seen few alternatives to the big tech players. *Data trusts* – "a legal structure that provides independent stewardship of data" – are proposed as one solution by the Open Data Institute [5]. They also define a spectrum of openness, ranging from closed data, via 'club' data shared with access control, to fully open data [5]. This is a potential structure for gradually maturing a community towards open data collaboration.

6 Threats to validity

Regarding external validity, focus group participants were selected using convenience sampling [20], and thus statistical generalization is not an option. However, for this exploratory, qualitative study, the primary focus is on diversity, which we report in Tables 1 and 2. Hence, our results are relevant although we cannot say which are more important than others. However, we might have overlooked some domain where ODC is more established, which we have tried to mitigate by reviewing related literature and reach out in broad, multi-domain industry networks.

A more significant threat is that we explore a topic (ODC) that is still in its infancy. Thus we collect opinions and hypotheses, rather than facts and experiences in relation to the ODC. Further, as the constructs are not well defined, we might misinterpret the participants. In order to mitigate threats to construct validity, we gave a short introduction to the ODC concepts in the beginning of each workshop. Further, among the participants, significant experience with OSS was represented, which gave a frame for the open concept.

On internal validity, the coding was performed by the second author who never had the secretary role. The first author had the secretary role for FG3. This procedure addresses researcher bias, as the codes are based on the notes by someone else and latter the codes are reviewed by the first author. Furthermore, we validated our preliminary results at the public event, which confirmed our conclusions broadly. This further addresses confirmation bias, even though this is still a potential threat to the validity of our work.

7 Conclusion and future work

We report the in-depth analysis of five focus group meetings on the concepts of Open Data Collaboration (ODC). Collecting input from 27 participants from 22 different private companies and public authorities, representing a variety of industrial and societal domains, provides a rich view of challenges and opportunities for ODC. We still believe that ODC will be one way to realize open innovation, both in public–private partnerships, but also between multiple private actors.

Open Innovation, whether through OSS, ODC, or other mechanisms, 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 may 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.

Being a concept in its infancy, ODC has to be further explored. As it involves an interplay with technical, organizational, legal, and business factors, we believe that these issues must be studied in pilot studies of practice in some sort of sand-boxing environment, technically and organizationally. We therefore work on setting up semi-open data collaboration. Currently we work on automotive and Industry 4.0 applications. Thereby, we aim to validate which challenges need to be handled, how data collaboration can be initiated and grow sustainably.

Acknowledgements

We thank our collaborator Sofie Westerdahl of Mobile Heights for co-organizing the workshops. Thanks to the participants in the focus groups for their contributions. Thanks also to Dr. Markus Borg, RISE, for reviewing an earlier version of this paper. This work was funded by the Swedish National Innovation Agency, VINNOVA, under grant 2018-04341 for groundbreaking ideas in industrial development.

References

[1] H. Chesbrough, W. Vanhaverbeke, and J. West, *New frontiers in open innovation*. Oup Oxford, 2014.

[2] 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.

[3] S. Jansen, S. Brinkkemper, J. Souer, and L. Luinenburg, "Shades of gray: Opening up a software producing organization with the open software enterprise model," *Journal of Systems and Software*, vol. 85, no. 7, pp. 1495–1510, 2012.

[4] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International journal of information management*, vol. 35, no. 2, pp. 137–144, 2015.

[5] D. Coyle, S. Diepeveen, and J. Wdowin, "The value of data summary report," The Bennett Institute, Cambridge, Tech. Rep., 2020.

[6] A. Raj, J. Bosch, H. Holmström Olsson, A. Arpteg, and B. Brinne, "Data management challenges for deep learning," in *45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2019, pp. 140–147.

[7] P. Runeson, "Open collaborative data - using OSS principles to share data in SW engineering," in *International Conference on Software Engineering (New Ideas and Emerging Research)*. IEEE / ACM, 2019, pp. 25–28.

[8] D. Rudmark and A. H. Jordanius, "Harnessing digital ecosystems through open data – diagnosing the Swedish public transport industry," in *Proceedings of the 27th European Conference on Information Systems (ECIS), Research-in-Progress Papers*, 2019.

[9] H. W. Chesbrough, *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press, 2003.

[10] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Information Quarterly*, vol. 32, no. 4, pp. 399–418, 2015.

[11] 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.

[12] S. S. Dawes, L. Vidiasova, and O. Parkhimovich, "Planning and designing open government data programs: An ecosystem approach," *Government Information Quarterly*, vol. 33, no. 1, pp. 15–27, 2016.

[13] T. Olsson and P. Runeson, "Open data collaborations: a snapshot of an emerging practice," in *Proceedings of the 15th International Symposium on Open Collaboration*. ACM, 2019, pp. 1–4.

[14] T. Olsson, P. Runeson, and S. Westerdahl, "Open collaborative data: a pre-study on an emerging practice," RISE, Tech. Rep., 2019. [Online]. Available: http://ri.diva-portal.org/smash/record.jsf?pid=diva2:1343126

[15] C. Alves, J. Oliveira, and S. Jansen, "Understanding Governance Mechanisms and Health in Software Ecosystems: A Systematic Literature Review," in *Enterprise Information Systems*, S. Hammoudi, M. Śmiałek, O. Camp, and J. Filipe, Eds. Springer, 2018, pp. 517–542.

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

[17] A. Zuiderwijk, M. Janssen, and C. Davis, "Innovation with open data: Essential elements of open data ecosystems," *Information Polity*, vol. 19, no. 1, 2, pp. 17–33, 2014.

[18] M. Kassen, "Understanding transparency of government from a nordic perspective: open government and open data movement as a multidimensional collaborative phenomenon in Sweden," *Journal of Global Information Technology Management*, vol. 20, no. 4, pp. 236–275, 2017.

[19] E. Styrin, L. F. Luna-Reyes, and T. M. Harrison, "Open data ecosystems: an international comparison," *Transforming Government: People, Process and Policy*, vol. 11, no. 1, pp. 132–156, 2017.

[20] C. Robson, *Real World Research*, 2nd ed. Blackwell, 2002.

[21] J. Kontio, J. Bragge, and L. Lehtola, "The focus group method as an empirical tool in software engineering," in *Guide to Advanced Empirical Software Engineering*, F. Shull, J. Singer, and D. I. K. Sjøberg, Eds. London: Springer, 2008, pp. 93–116.

A Focus group guide

Below is a list of topics and questions to guide the workshops. They should not all be answered, rather it is kind of a checklist. For each of the three sections, spend 5 min individually on post-it notes, 10 minutes presentation in group, and 10 minutes discussion.

A.1 Individual notes

What types of data does your company collect/handle/use? Examples?

A.2 Characteristics of data collection

1. Are data collected as input to the development of the product/services or to the continued operation?
2. What is the lead-time from a phenomenon occurs to that it can be observed in collected data? What is the lifetime of data?
3. How much effort needs to be put into processing the data before it is possible to make analysis?
4. To what extent is the analysis automatic?
5. To what extent are privacy issues related to the data collected?
6. Are you using ML today? Big data?

A.3 Individual notes

Which data can be shared? Under which conditions? To whom?

A.4 Sharing data

1. Is the data shared with other organizations?
2. Is the data a competitive advantage? Same domain/different domains?
3. Can it be a differentiator to share data – and thereby being part of a community or ecosystem?
4. What are the costs related to collecting data?
5. How unique is the data to your organization?
6. What would happen if you stop collecting data?
7. Are you charging others to the data you are sharing to them?
8. Do you make data publicly available without charge or other (direct) monetary incentives? Altruistic?

A.5 Bridges and barriers

1. Technical challenges in collecting data? Sharing data? – Cloud, connectivity, bandwidth, security, etc.
2. Legal barriers – GDPR
3. Business – Competition, differentiation
4. Have you had security incidents where unauthorized individuals have gotten access to data? What type of data was accessed?
5. Authenticity – How do you ensure the data is authentic?
6. Costs – Can cooperation reduce the cost of data collection?

A.6 Prepare for presentation

Prepare summary in three slides according to sections above.