Open Data Collaborations – a snapshot of an emerging practice

Thomas Olsson
RISE Research Institutes of Sweden
Scheelevägen 17
223 70 Lund, Sweden
thomas.olsson@ri.se

Per Runeson
Lund University
Ole Römers väg 3
221 00 Lund, Sweden
per.runeson@cs.lth.se

 ACM ISBN 978-1-4503-6319-8/19/08.
https://doi.org/10.1145/3306446.3340832

Abstract
Data defined software is becoming more and more prevalent, especially with the advent of machine learning and artificial intelligence. With data defined systems come both challenges – to continue to collect and maintain quality data – and opportunities – open innovation by sharing with others. We propose Open Data Collaboration (ODC) to describe pecuniary and non-pecuniary sharing of open data, similar to Open Source Software. To understand challenges and opportunities with ODC, we ran focus groups with 22 companies and organizations. We observed an interest in the subject, but we conclude that the overall maturity is low and ODC is rare.

Author Keywords
Open data, Data collaborations, Open innovation

ACM Classification Keywords
D.2.0 [SOFTWARE ENGINEERING]: .

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 of machine

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

learning and artificial intelligence (ML/AI). Components, which behavior are defined by data are more and more common in software systems. Access to data is often 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 advantages anymore. Still, data has to remain reliable and of high quality to be useful for the companies, which implies costs for maintenance and quality assurance. We propose the concept of ODC – Open Data Collaboration – as a means for managing collaboration and to cooperate across companies on open data [6].

Open Source Software (OSS) is utilized in almost any software system, including 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" [2]. 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 [5].

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 organized a series of focus groups with practitioners. In the focus groups, participants from different companies discussed issues related to data, open data and open data collaborations.

For this poster, we highlight the concepts around ODC and the preliminary results of the focus groups held with companies and public organizations in Sweden.

Related work

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. The exchange of data is inevitable.

There are initiatives on sharing data for machine learning, like Open Mined[2]. 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.

Frizzo-Barker et al. [4] conducted a systematic mapping study of research on Big Data in business scholarship. With respect to openness, they identify open collection of data (crowdsourcing) and open tools for big data analysis.

Del Vecchio et al. [3] provide an extensive analysis of research on the borderline between information systems and innovation management, with focus on open innovation. Open source platforms, such as Hadoop, contributes to open innovation based on Big Data, and analysis of data may lead to business innovation.

Focus groups on Open Data Collaborations

To understand the challenges and opportunities facing companies around ODC, we organized several three-hour focus group workshops. Our goal with the focus groups was to achieve an interaction and experience sharing

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

among a diverse group of companies and public organizations, and thus understand more of the challenges and opportunities of ODC.

Each of the focus group sessions followed a similar scheme. First, an introduction to the concepts of ODC were given. Then, the attendants were split into focus groups of 5-7 participants, plus one moderator and one secretary.

The composition of the focus groups was made based on the affiliation of the participants; company, research organization or public agency, to create a variety of participants in the discussion. Finally, the groups reassembled, and a summary of each group was presented and discussed.

We sent out open invitations where any company and organization could register to participate. We grouped participants to allow for meaningful interaction but avoided to have participants from the same organization in the same group. In total, we organized three workshops with 27 participants from 22 companies, public organizations and non-profit organizations. In two of the instances, we split into two focus groups, thus running in total five focus groups.

The collected data from each of the five focus groups was quickly summarized in the follow-up session, based on the three main themes. We plan to conduct a more thorough analysis to find nuances in the findings below.

Preliminary results

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.

Data is not always possible to purchase
Certain types of data can be purchased, such as market data and data collected by smart phones and apps like Facebook. Marketplaces and data brokers exist for this. Furthermore, there are open initiatives, e.g. openmined, to share open 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.

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 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.

There are costs of working with and for sharing data to consider
To collect useful data and ensuring its quality for the intended purpose 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, to ensure reliable and secure communication as well as additional mechanisms to filter out which data to actually share.

Quality and trust are key issue
As data becomes more and more important for successful development and reliable operation, the requirements on quality for 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.

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.

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 also integrity issues, specifically in the light of GDPR.

Conclusion and future work

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 overall insight that open data collaborations will be needed was largely lacking. From the study, we identify four areas of future research:

1. Technical – sharing of data requires an infrastructure and mechanisms.
2. Organizational – processes and competencies have to be in place.
3. Business – there is a lack of models to estimate value as well as costs.
4. Legal – There are legal aspects around data whether sharing or receiving data.

We are currently discussing project ideas with different companies to deepen our understanding of ODC for specific application domains.

References

[1] Axelsson, J. Systems-of-systems for border-crossing innovation in the digitized society-a strategic research and innovation agenda for sweden, 2015.

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

[3] Del Vecchio, P., Di Minin, A., Petruzzelli, A. M., Panniello, U., and Pirri, S. Big data for open innovation in smes and large corporations: Trends, opportunities, and challenges. *Creativity and Innovation Management 27*, 1 (2018), 6–22.

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

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

[6] Runeson, P. Open collaborative data – using oss principles to share data in sw engineering. In *ICSE (NIER)*, IEEE / ACM (2019), 25–28.