2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)

Open Collaborative Data – using OSS principles to share data in SW engineering

Per Runeson
Dept. Computer Science, Lund University,
Lund, Sweden
Email: per.runeson@cs.lth.se

***Abstract*—Reliance on data for software systems engineering is increasing, e.g., to train machine learning applications. We foresee increasing costs for data collection and maintenance, leading to the risk of development budgets eaten up by commodity features, thus leaving little resources for differentiation and innovation. We therefore propose Open Collaborative Data (OCD) – a concept analogous to Open Source Software (OSS) – as a means to share data. In contrast to Open Data (OD), which e.g., governmental agencies provide to catalyze innovation, OCD is shared in open collaboration between commercial organizations, similar to OSS. To achieve this, there is a need for technical infrastructure (e.g., tools for version and access control), licence models, and governance models, all of which have to be tailored for data. However, as data may be sensitive for privacy, anonymization and obfuscation of data is also a research challenge. In this paper, we define the concept of Open Collaborative Data, demonstrate it by map data and image recognition examples, and outline a research agenda for OCD in software engineering as a basis for more efficient evolution of software systems.**

I. Introduction

"Data is the new oil" is a mantra coined by Clive Humby[1] in 2006. In the last decade, an enormous amount of companies have grown based on Big Data. With the growing interest in machine learning, the data becomes an input to the engineering of software, in that the behavior of the software is defined and modified based on training data over time.

"Software is eating the world" is another mantra, this coined by Marc Andreessen in 2011 [1]. One aspect of this statement is that software is eating the R&D budget of companies, especially commodity software. As a consequence, less is left for differentiating features, as identified by Bosch [2] in his three layer product model (commoditized, differentiating, and innovation layers). To stay competitive, Bosch advices that companies should *"make a clear distinction between the layers and thus to allocate resources appropriately to encourage development at the upper* [innovation] *layer."*

Combining these two trends, i) the growing reliance on data, and ii) increasing costs for software maintenance, leads us to claim that *data also adheres to the three layer model of commodity, differentiation and innovation*, and that *costs for data maintenance is an upcoming challenge for software companies*. The cost of curating and maintaining data, will sooner or later exceed its business value.

One approach to adress this issue in software, is open sourcing what has no or little differentiating value anymore. Thereby, the maintenance costs may be shared by multiple companies using the commodity software. As a results, more differentiation can be achieved, and other positive side effects of open innovation may be gained [3], i.e., inflow of ideas and knowledge for innovation. In fact, studies show that the inflow of innovation may be the dominating gain even if the open sourcing was initiated to save costs [4].

We therefore propose *Open Collaborative Data (OCD)* as an approach to share data and reduce maintenance costs for software companies, under kept or improved competitiveness, similar to open source software (OSS). Paraphrasing a definition of OSS [5, p.5], *Open Collaborative Data (OCD) is a type of digitally stored data which is released under a license in which the copyright holder grants users the rights to study, process, and distribute the data to anyone and for any purpose.*

While OCD is inspired by OSS, software engineering research has to adress several technical and managerial challenges in relation to OCD, where it differs from the OSS counterpart. There are initiatives, like Open Knowledge Foundation[2], which provide guidelines for sharing data, but research surveys conclude there is no systematic research on data sharing in software engineering.

We present background work on open innovation, open source software, and open data, underpinning our claim in Section II. Section III presents two examples of data sharing, and Section IV sketches a research agenda based on the identified needs. Section V concludes the paper.

II. Background – commercial openness and data

OSS in commercial business has emerged as a means 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 BSD[3]. Studies on open software tools [4] as well as on product

[1]UK mathematician and architect of Tesco's Clubcard

[2]https://okfn.org/opendata/

[3]https://en.wikipedia.org/wiki/History_of_free_and_open-source_software

978-1-7281-1758-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICSE-NIER.2019.00015

software [6] indicate that OSS plays a key role for software business, although it has to be managed accordingly.

Chesbrough coined the term Open innovation (OI) [3], 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"* [7]. 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 organisational boundaries.

Open Data is brought forward as an enabler for innovation and entrepreneurship, e.g., by Lakomaa and Kallberg [9]. 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 [10]. 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]. For commercial companies, Fabijan *et al.* identify problems with sharing data even *within* an organisation [11], and we have not found any research on companies sharing data *between* organizations.

Federated learning is a research branch of machine learning, where models are trained on multiple sets of data in their own context. Hence, the training model is brought to the data, rather than the data to the model. There are recent applications, sharing electronic health records, without revealing their sensitive content [12]. While federated learning addresses privacy concerns related to data sharing, it introduces threats to the transparency in the learning process, which is another key aspect of automated decision making.

Frizzo-Barker *et al.* [13] map research on Big Data in business scholarship. With respect to openness, they only identify open collection of data (crowdsourcing) and OSS tools for big data analysis. Other challenges include how to i) take advantage of the enormous volumes of data, ii) handle the risk of privacy and ethical infringements, and iii) manage 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.

Del Vecchio *et al.* [14] provide an extensive overview and analysis of research on the borderline between information systems and innovation management, with focus on OI. They report how OSS platforms, such as Hadoop, contribute to OI 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., while research on sharing data between corporations as a means to foster OI was absent.

III. Scenarios on Open Collaborative Data

Given these findings in the literature on open software, open data, and open innovation, combined with the fact that the importance of data is growing for several types of applications, lead us to think of OCD as a potential means for spending less on commodity features and more on differentiating and innovative features in data-driven applications. We demonstrate the case with two example scenarios and draw from them on the generality of OCD.

A. Map based applications

Several applications – mobile apps as well as business applications – depend on maps. Examples include navigation services, but also analysis tools for transportation, analysis of business or governmental activities, like spatial planning, etc. For most of these applications, the map is not a differentiating feature, but a nescessity for any application, i.e., part of the commodity. Users do not choose one app before the other based on the map quality, which they did, e.g., when Apple and Google were competing about having the best map data[4]. The map data is reasonably stable for the largest part, while all stakeholders are interested in continuously ensuring the quality of the data and quickly getting updates according to changes. For example, new or closed roads should be incorporated in the map data and faults be corrected. In that sense, map data is to a large extent commodity for a developer of map based applications.

There is an OCD initiative, Open Street Map[5], providing open map data, which shows the feasibility of open collaborative data approaches. The community is governed similarly to an OSS community, although we hypothesize that there is more to learn from comparing it to OSS working and governance practices, and that other types of data may be shared and governed similarly.

B. Image recognition

In the field of image recognition, machine learning approaches have grown significantly, to provide various kinds of services. A delimiting factor for many applications is the access to labeled training data. This may lead to biased algorithms, as in the case of face recognition of minorities [15].

There are databases available for face recognition research purposes[6]. However, for commercial applications, there does not seem to be any clear model for sharing or monetizing image data for machine learning purposes. Further, the introduction of new legislation to protect privacy and increase transparency in decision making (such as GDPR[7] in Europe)

[4]See e.g. a discussion in Forbes https://www.forbes.com/sites/quora/2017/06/08/google-wouldnt-negotiate-with-apple-to-keep-maps-on-ios-devices-and-that-was-the-wrong-move/#1bfad14763a9

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

[6]See e.g., https://skymind.ai/wiki/open-datasets

[7]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

Table I
COMPARISON BETWEEN OCD EXAMPLES OF MAP AND IMAGE DATA, AND OSS

	Map data	Image data	OSS
Technical infrastructure	Specialized web tools	Specific schemes for each data set	General, mature open tools
Licence model	Open database licence	For research only, sometimes CC	Several established licence models
Governance	Run by foundation or corporation	No dominating model	Cathedral and bazaar, and all in between
Privacy	Sensitive data about objects	Sensitive data about people	Contributor identity

adds to the demands and costs for collecting, storing and sharing data, for machine learning applications.

Going beyond pure image recognition, into scenario recognition in videos, e.g., for autonomous driving, requires even more data for learning. This is identified as major obstacle, and leads development towards simulating scenarios, which are used as learning inputs to the autonomous cars[8].

Today, this annotated video data is a differentiating asset for autonomous car manufacturers, but gradually, that will also turn into commodity data, and companies have to find ways to reduce the data costs to release funds for development of new innovations.

C. Summary

The two examples indicate that there are communities already sharing data in a way that resembles OSS. Other potential areas of OCD include transportation (information about parking lots, congestions, road work etc.), weather and climate (sensing information from various local climate control systems), healthcare (sensing information, tests, longitudinal series), culture and tourism (pictures, reviews) etc. However, an OCD community is different from an OSS community, although we have not seen any systematic analysis of that difference. Below, we analyze the two examples and compare them to general OSS practices.

IV. RESEARCH AGENDA FOR OPEN COLLABORATIVE DATA

To identify what is needed in terms of software engineering research, to make OCD feasible for companies in their software engineering process, we compare the open data examples above to OSS practices. We are aware of OSS practices not being homogeneous. However, we still argue the comparison is relevant in order to identify a research agenda for OCD. The comparison is summarized in Table I and elaborated below.

The *technical infrastructure*, based on the internet in general, and particularly on collaboration and configuration management tools, is one of the cornerstones for OSS [16]. The development of, e.g., Git as a distributed configuration management tool, Jira for issue management, and Slack for communication, have contributed significantly to making OSS widely adopted. In contrast, Open Street Map has specific tools, tailored for the application, to collect data from users. They offer APIs and download options to fit different needs. The image recognition open datasets are based on established standards, such as .jpg, but beyond that, each dataset is stored and labeled according to specific coding schemes.

Licence models for OSS is a complex area, with different licence models emerging to balance the needs for corporations to keep some code open (commodity), while protecting other code, which is their competitive advantage and innovation base. For their map data, Open Street Map uses a specialized open data licence, Open Data Commons' Open Database License (ODbL), which is an attribution and share-alike license for data and databases[9], while some documents use the Creative Commons (CC) framework[10]. The image databases we have found use different kinds of "for research only" licences, sometimes based on CC.

Governance in the open source context refers both to the community internal and the proprietary vs. open aspects. Raymond's classical The Cathedral and the Bazaar [17] addresses the first aspects. The cathedral represents a controlled community, with an exclusive group of contributors, while the bazaar model represents the open culture of a multitude of collaborators. The other governance dimension, on how the OSS and the pr oprietary can coexist, interfaces towards law and innovation management research, e.g by Kemp [18] and Munir *et al.* [4]. The map data example above demonstrates the most open option for proprietary businesses, where the Open Street Map foundation is open for membership, and all data is shared free of charge. Other governance models exist, e.g., Mapillary's[11], where images are made open in relation to the Open Street Map, while their additional identification service is payment based. In the field of image recognition, we are not aware of any dominating commercial actor, but assume that there are bi-corporate agreements on data sharing.

In OSS, *privacy* is primarily related to the data contributors, whether or not revealing their name and identity, and thus making contributions traceable to individuals. For OCD, the privacy issue is much more complex, and is related to individuals or phenomena in the data as such. Sensitive information may be personal images, but also positions of classified buildings and infrastructure.

Based on the observed differences between OSS and

[8]https://www.theverge.com/transportation/2018/4/19/17204044/tesla-waymo-self-driving-car-data-simulation

[9]https://opendatacommons.org/licenses/odbl/

[10]https://creativecommons.org

[11]https://www.mapillary.com

OCD, summarized in Table I, we propose a research agenda on OCD to be adressed in software engineering:

- Technical infrastructure – what general technical infrastructure is needed to support OCD? What kind of collaboration takes place around OCD? What tool support is needed to protect privacy and ensure compliance with legal and ethical standards?
- Data licence models – which data licence models are needed for OCD? Do general models, such as ODbL apply to different kinds of data, or do different data require different licences?
- Data governance – how can data be governed? Which restrictions of usage do contributors want to set, and how does that change of time (e.g., via the GDPR "right to be forgotten")?
- Privacy concerns – how can data be shared without infringing privacy rights, e.g., through data obfuscation or anonymization? Can federated learning approaches be made more transparent?

We do not claim this list is complete, but rather an indication, based on the example cases, that there is a need for research on how to collaborate around shared data, in an OSS-like fashion.

V. Conclusions

We identify that data plays an increasingly important role in software engineering, for machine learning applications and for data in applications. We hypothesize that Open Collaborative Data (OCD) will help adress these needs for commodity data, implying that development resources may be saved for differentiation and innovation. Although OCD resembles OSS, there are several issues which are different, especially with respect to the privacy concerns with data. We therefore propose OCD become a topic of software engineering research, exploring and guiding aspects of technical infrastructure, data licence models, data governance, and privacy. Well managed, OCD can contribute, not only to being "the new oil", but "the renewable energy" in an open, collaborative innovation context.

Acknowledgement

Thanks to Thomas Olsson, RISE, for inspiring discussion on this and related topics, and feedback on an earlier version.

References

[1] M. Andreessen, "Essay why software is eating the world," *The Wall Street Journal*, August 20 2011.

[2] J. Bosch, "Achieving simplicity with the three-layer product model," *Computer*, vol. 46, no. 11, pp. 34–39, 2013.

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

[4] 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.

[5] A. M. St. Laurent, *Understanding Open Source and Free Software Licensing*. O'Reilly Media, 2008.

[6] 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.

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

[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] 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.

[10] 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.

[11] A. Fabijan, H. H. Olsson, and J. Bosch, "The lack of sharing of customer data in large software organizations: Challenges and implications," in *Agile Processes, in SE, and XP*, H. Sharp and T. Hall, Eds. Springer, 2016, pp. 39–52.

[12] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records." *International Journal of Medical Informatics*, vol. 112, pp. 59 – 67, 2018.

[13] 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*, vol. 36, no. 3, pp. 403 – 413, 2016.

[14] 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 Innov. Mgmt*, vol. 27, no. 1, pp. 6–22, 2017.

[15] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conference on Fairness, Accountability and Transparency*, ser. Proc. of ML Research, S. A. Friedler and C. Wilson, Eds., vol. 81, New York, USA, 2018, pp. 77–91.

[16] A. Mockus, R. T. Fielding, and J. D. Herbsleb, "Two case studies of open source software development: Apache and mozilla," *ACM Trans. Softw. Eng. Methodol.*, vol. 11, no. 3, pp. 309–346, Jul. 2002.

[17] E. S. Raymond, *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly Media, 1999.

[18] R. Kemp, "Open source software (OSS) governance in the organisation," *Computer Law & Security Review*, vol. 26, no. 3, pp. 309 – 316, 2010.