Collaborative Aspects of Open Data in Software Engineering

Linaker, Johan; Runeson, Per; Zuiderwijk-van Eijk, A.M.G.; Brock, Amanda

DOI
10.1109/MS.2021.3118123

Publication date
2022

Document Version
Final published version

Published in
IEEE Software

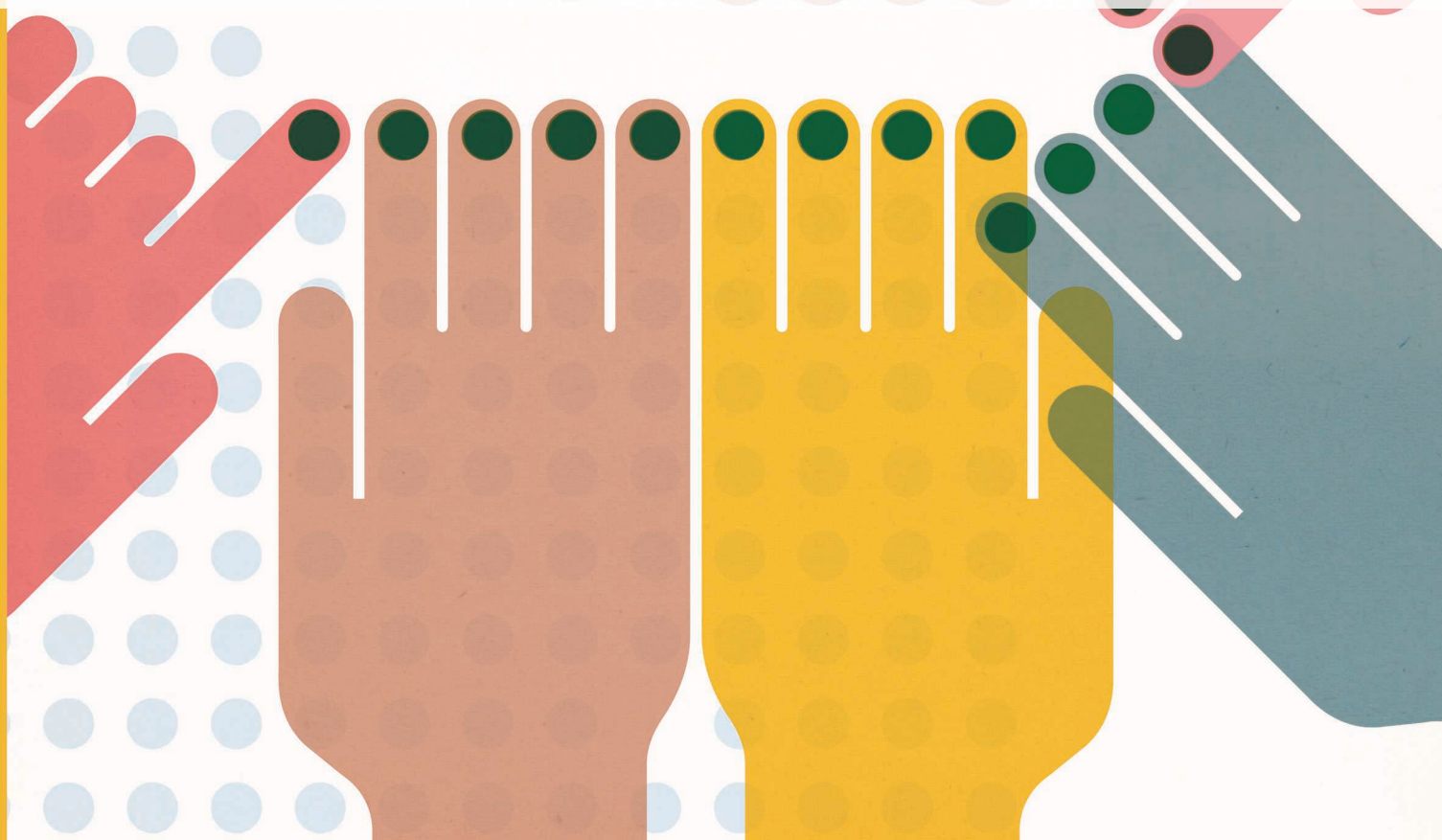
Collaborative Aspects of Open Data in Software Engineering

Johan Linåker, RISE Research Institutes of Sweden

Per Runeson, Lund University

Anneke Zuiderwijk, Delft University of Technology

Amanda Brock, OpenUK



Digital Object Identifier 10.1109/MS.2021.3118123
Date of current version: 23 December 2021

ENGINEERS REQUIRE HIGH-quality data for the design and implementation of today's software, especially in the context of machine learning (ML). This puts an emphasis on the need for the publication and sharing of data from and between organizations, public as well as private. Following the paradigm of open innovation, open data provide a mechanism to increase the availability of information, offering utility and improving innovation and user choice through the inevitable interoperability this enables.

Consistent with previous work, this editorial defines *open data* as machine-readable information proactively shared on the Internet under a license that gives people the right to use, process, and distribute the material to anyone and for any purpose.[1] The *open* in *open data* highlights the potential for innovation and collaboration, which can range from sharing new and connected data sets (including metadata) to quality assurance, or the enrichment and life cycle management of information. Collaboration facilitates related standards and formats for sharing and use [commonly in the form of open source software (OSS)], and it extends to the platform (i.e., software and infrastructure) that is employed.

However, collaboration through openly sharing data sets is also complex and challenging. For example, it may be unclear which data sets are used for software and application development. Furthermore, open government data have barely been studied from a software development perspective. As a consequence, there is a lack of insight into the requirements of the software development community and how it leverages open data,[2] although an understanding has begun to emerge in the literature.[3] This special issue of *IEEE Software* focuses on the collaborative aspects of open data in software engineering.

Collaborative Aspects of Open Data in Software Engineering

Collaboration on open data often occurs in what is called *open data ecosystems* (*ODEs*), where OpenStreetMap and Wikidata are two popular examples. These ecosystems represent a form of networked communities of actors (organizations and individuals) that base their relationships with one another on common interests.[4] Common interests create and provide free geographic data in the case of OpenStreetMap and a free knowledge base that can be read and edited by humans and machines in the case of Wikidata.

An ODE is commonly supported by a technological platform that enables actors to process data (e.g., find, archive, publish, consume, and reuse) and foster innovation, create value, and support new businesses. Actors collaborate on data and boundary resources (e.g., software and standards) through the exchange of information and artifacts. In the cases of OpenStreetMap and Wikidata, their respective platforms are openly available, including boundary resources, such as software for publishing and managing data.

Another essential characteristic of an ODE is how its governance is established, and we differentiate between organization-centric, consortium-based, and community-based approaches.[4] OpenStreetMap and Wikidata are community based, where governance is spread among members of the ecosystem. In organization-centric and consortium-based ODEs, governance is concentrated with a single actor or a set group of actors, which are commonly public or private organizations with similar business interests in the data shared within the ecosystem. Examples from research can be found within domains such as the labor market, public transport, smart cities, and Industry 4.0, where public and private organizations hold and share governance in different constellations.[3]

Several actors may be involved in the collaboration in an ODE, creating a value chain ranging from data providers to information users. Lindman et al.[5] identified five roles in this type of collaboration: open data publishers, data extractors and transformers, data analyzers, user experience providers, and support service providers. These are needed to get a fully functional pipeline from data to user service, and they may be filled by actors within or across organizations. A data publisher is commonly a public entity, as sharing open data from private parities is not yet a common phenomenon.[4]

As with all aspects of digitalization, open data involve collaboration between actors with competence in the digital domain and parties knowledgeable in the application domain. Successfully combining digital competences with application domains requires willingness and an ability to cross cultural and language barriers. Further, as legal conditions for the use and spreading of data are foundational for data-driven software development, an ability to understand and communicate with jurists is essential.

Benefits of Collaboration on Open Data

Collaboration on open data has potential to generate value similar to OSS and other types of open innovation. This comes as an effect of tapping the wisdom of the crowd and exploiting the potential workforce within and outside an ODE. From a cost-saving perspective, the potential external workforce may help with tasks such as providing, collecting, processing, and publishing data.[4] As highlighted, an ODE can resemble a value chain, where the raw material of data gets enriched and

processed in a collaborative manner.[5] Reaching outside an ODE, crowdsourcing as well as mass collaboration can be used to increase participation and teamwork among external individuals.[6]

From a quality perspective, collaboration on open data, e.g., can help to address and correct errors, and it can add information through annotations and other kinds of metadata.[4] As a consequence, the quality of ML training sets and software that use the data will be improved. This applies to open governmental data,[3] where users may help improve information quality for public agencies.

From an innovation point of view, the potential for increased access to high-quality data can help provide new and extended training sets for ML-based applications as well as feature sets for other forms of data-driven software (e.g., Google Maps). Innovation can be accelerated, as new use cases and markets may be extended or created for an ODE. Furthermore, collaborating through open data may lower entry costs for actors aiming to utilize data to offer services, help catalyze new entrepreneurial efforts, increase transparency and accountability, and transform incumbents and public organizations through improved decisions and services.[7]

Challenges and Ways of Improving Collaboration on Open Data

Sharing and collaborating on open data bring a number of challenges, technical and process-oriented as well as cultural and business-oriented.[4] In the following, we highlight a few of the challenges that practitioners may need to consider within an ODE or when thinking of entering or creating one

Business and Competition Aspects

From a business perspective, an important challenge concerns the motivation of why a data set should be shared in the first place.[3] There needs to be incentives that align with a company's business model. Practitioners thus need to understand the aforementioned benefits and be able to contextualize them in their own environment and relate them to relevant business goals.

The benefits need to be nuanced and weighed against the potential costs and risks of releasing the data. Costs may be related to the data management life cycle, i.e., the collection, processing, quality assurance, sharing, and distribution of the information. As with OSS, these costs as well as the potential benefits relate to the amount of collaboration that actually takes place. Hence, actors within an ODE must find ways for facilitating and orchestrating sustainable collaboration and sharing.

A specific challenge for such collaboration is the notion of coopetition,[4] i.e., a space where business rivals can work with one another without being afraid of giving away or losing their competitive edge,[8] which can be a significant obstacle to commercial data sharing. Researchers and companies may not be willing to openly share their data about novel software innovations and services since this reduces their ability to commercially exploit them.[9] On the other hand, commodity data may be a basis for coopetition, as, for example, OpenStreetMap demonstrates.

To manage and enable such coopetition, there may be a need for a neutral governance actor within the ecosystem that can mediate discussions, craft a common vision, and help actors share data that everyone is comfortable with.[3] The latter, commonly referred to as *selective revealing*, can mean, e.g., that only certain abstractions of data are shared.

Technical Aspects

The potential collaboration aspects also bring up a number of more technical challenges.[4] One concerns the collection of data and how to ensure their quality. Some scholars indicate that data sets of insufficient quality may be misinterpreted or misused,[10] rather than improving the quality of software. Common domain models and standards for how data are shared and used, as well as transparent processes and OSS tools for the collection and enrichment of information, address this challenge.

Introducing feedback loops within an ODE and toward end users is another means, which is specifically highlighted and addressed by Rudmark and Andersson in this special issue. Versioning data is another practical aspect that needs consideration to enable decentralized collaboration, similar to what can be observed in OSS ecosystems. Worthington et al. explore this topic further in their contribution.

Cultural, Organizational, and Legal Aspects

Cultural and organizational aspects form another set of challenges,[4] e.g., aligning strategic and operational levels within an organization about what data to share. Individuals may have different views and understandings of the risks and benefits that sharing would imply. These challenges may also manifest in the collection of and collaboration around information, as, again, individuals of different backgrounds and cultures may have unaligned perspectives. This is further explored by Thurnay et al., who highlight the need for practitioners to educate one another based on their different domain knowledge and understanding.

Another set of challenges relates to legal conditions.[4] Organizations

may be reluctant to share data due to uncertainty about liability and what licenses may imply in practice. The risk of legal complications due to the General Data Protection Regulation is a specific and common concern under European legislation. Companies, for example, collect and maintain considerable proprietary employee and customer data that they may be reluctant to share through collaborations.[5] Enabling individuals to gain control of how their data are shared (see https://mydata.org) may be one way of addressing this difficult challenge, an area that is further explored by Alorwu et al. in their contribution.

Overview of the Special Issue Articles

This special issue covers different topics related to collaborative aspects of open data in software engineering. Of the nine submitted papers, we selected four. We applied a rigorous review process to each article, including a review by at least three experts. We summarize the articles in the following.

The first one deals with originators of data. Alorwu et al. elaborate on the contributors of crowdsourced data in the health domain. Open health data may be used for research and as inputs for software solutions. The authors surveyed 80 participants who previously donated health data to a decision support system, asking about their willingness to donate information for public use. They find that donators, despite giving information as "open," wanted to influence what, by whom, and where their data were used. The respondents voice limited trust in private stakeholders, such as pharmaceutical and insurance companies, and raise privacy concerns. The authors conclude by connecting these concerns with the MyData initiative, providing mechanisms for donors to keep control of their information while still allowing it to be used for certain purposes.

Next, Rudmark and Andersson focus on the quality of data and the role feedback loops may have in improving it. Feedback may be given by data publishers themselves and by data users. Further, feedback may come from internal and external uses. With examples from public transport information, the authors present data dogfooding, where providers use their own information; external application monitoring, where providers monitor how their data are used in other actors' applications; community curation, where actors work together on improving data quality; and external quality proxies, which involves letting an external actor check data before publishing them. These approaches are applicable depending on the characteristics of a data ecosystem.

The need for data users to understand what has changed is addressed by Worthington et al. They report a case involving open customs tariff data and observe users' strategies to overcome a lack of change information. Based on their observations, the authors outline three approaches to communicate changes in open data sets: 1) publish change information as a "sidecar," i.e., outside the core system as a separate information entity, 2) publish versions with change tracks as an HTTP application programming interface end point to be consumed by users, and 3) integrate versioning into databases and allow users to query changes. The case demonstrates how important it is that open data not only are published but set under version control.

In the fourth article, Thurnay et al. address the cross-discipline communication and collaboration needed to create an ODE. They created a database of legal documents in Austria and report their lessons from collaboration between law experts and technology experts. They were surprised by how much implicit domain knowledge each category expert had and how much effort it took to bridge the gap. However, they took the role of teacher for one another, and the technology experts gradually became genuine experts in the narrow field of law, while the legal experts were able to read Python source code implementations of text processing.

Summary of Practice

Demand for high-quality data is growing as are the costs and resources for collecting and managing information. As a response, open data offer a new arena for collaboration and innovation, similar to OSS. Yet, software engineering practices have not kept pace in terms of enabling such collaboration. In part, there is a need for the evolution of ODEs, similar to open source communities, that can facilitate collaboration and data sharing among actors, independent of aligning and competing interests.

Challenges include technical, cultural, organizational, and legal aspects, as explored by the four articles in this special issue. Together, the articles clearly show that creating ODEs is a sociotechnical endeavor, and they provide practice-based recommendations to mitigate problems that might appear. The request for high-quality information for a multitude of data-driven applications will not decline, and open data crowdsourcing, quality assurance, version control, and collaboration will continue playing a central role in the future of software engineering.

Acknowledgments

We are incredibly grateful for the support we received from several individuals who made this special issue possible and successful, including the

anonymous reviewers, authors, and *IEEE Software*'s editorial staff.

References

1. "Open definition version 2.0," Open Knowledge Foundation. [Online]. Available: http://opendefinition.org/od/
2. M. Grzenda and J. Legierski, "Towards increased understanding of open data use for software development," *Inform. Syst. Front.*, vol. 23, no. 2, pp. 495–513, 2021. doi: 10.1007/s10796-019-09954-6.
3. J. Linåker and P. Runeson, "How to enable collaboration in open government data ecosystems: A public platform provider's perspective," *JeDEM—eJ. eDemocracy Open Government*, vol. 13, no. 1, pp. 1–30, 2021. doi: 10.29379/jedem.v13i1.634.
4. P. Runeson, T. Olsson, and J. Linåker, "Open data ecosystems–An empirical investigation into an emerging industry collaboration concept," *J. Syst. Softw.*, vol. 182, p. 111,088, 2021. doi: 10.1016/j.jss.2021.111088.
5. J. Lindman, T. Kinnari, and M. Rossi, "Business roles in the emerging open-data ecosystem," *IEEE Softw.*, vol. 33, no. 5, pp. 54–59, 2015. doi: 10.1109/MS.2015.25.
6. S. Takagi, "Research note: An introduction to the economic analysis of open data," *Rev. Socionetwork Strategies*, vol. 8, no. 2, pp. 119–128, 2014. doi: 10.1007/s12626-014-0048-6.
7. J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Government Inform. Quart.*, vol. 32, no. 4, pp. 399–418, 2015. doi: 10.1016/j.giq.2015.07.006.
8. H. A. Piwowar, R. S. Day, and D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate," *PLoS ONE*, vol. 2, no. 3, pp. 1–5, 2007. doi: 10.1371/journal.pone.0000308.
9. A. Zuiderwijk, R. Shinde, and W. Jeng, "What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption," *PLoS ONE*, vol. 15, no. 9, p. e0239283, 2020. doi: 10.1371/journal.pone.0239283.
10. A. Zuiderwijk and M. Janssen, "The negative effects of open government data—Investigating the dark side of open data," presented at the Proc. 15th Annu. Int. Conf. Digital Government Research, Aguascalientes, Mexico, June 18–21, 2014.

ABOUT THE AUTHORS

JOHAN LINÅKER is a senior researcher at RISE Research Institutes of Sweden, Lund, 211 00, Sweden. His research interests include empirical software engineering in industry and the public sector in the context of open source software, open data, and other fields of open innovation. Linåker received a Ph.D. from Lund University. Contact him at johan.linaker@ri.se.

PER RUNESON is a professor of software engineering at Lund University, Lund, 211 00, Sweden, where he leads the Software Engineering Research Group. His research interests include empirical investigation and collaboration with industry on software development and management methods. He serves on the editorial board of *IEEE Transactions on Software Engineering, Software Testing, Verification, and Reliability*, and the advisory board of *Empirical Software Engineering*. Contact him at per.runeson@cs.lth.se.

ANNEKE ZUIDERWIJK is an assistant professor in the Faculty of Technology, Policy, and Management, Delft University of Technology, Delft, 2628BX, The Netherlands. Her research interests include open data, specifically, theory development concerning infrastructural and institutional arrangements that incentivize information sharing and use by governments, researchers, companies, and citizens. Zuiderwijk received a Ph.D. from Delft University of Technology. She is the editor in chief of *e-Journal of e-Democracy and Open Government*. Contact her at a.m.g.zuiderwijk-vaneijk@tudelft.nl.

AMANDA BROCK is the chief executive officer of OpenUK, Rugby, CV21 3DE, U.K. Brock received an LLM in the fields of IP and IT law from from Queen Mary University of London. Contact her at amanda.brock@openuk.uk.