Challenges to Open Collaborative Data Engineering

Philip Heltweg
Friedrich-Alexander-Universität
Erlangen-Nürnberg
philip@heltweg.org

Dirk Riehle
Friedrich-Alexander-Universität
Erlangen-Nürnberg
dirk@riehle.org

Abstract

Open data is data that can be used, modified, and passed on, for free, similar to open-source software. Unlike open-source, however, there is little collaboration in open data engineering. We perform a systematic literature review of collaboration systems in open data, specifically for data engineering by users, taking place after data has been made available as open data. The results show that open data users perform a wide range of activities to acquire, understand, process and maintain data for their projects without established best practices or standardized tools for open collaboration. We identify and discuss technical, community, and process challenges to collaboration in data engineering for open data.

1. Introduction

Open data can be created, used, modified, and shared by anyone and therefore has the potential to be a driver of innovation. Still, research has shown that users face challenges when using open data. Part of these challenges are technical issues that make accessing the data hard. Additionally, the quality of data sources is often poor (Purwanto et al., 2020).

Data engineering, the activity and process of preparing data for use for a specific purpose, is costly and routinely consumes large parts of the budget for data science projects (Terrizzano et al., 2015). The importance of data engineering in open data is even larger because of the varying quality of data provided by publishers.

However, because open data can be freely modified and distributed, a community of users can share the work of making data easier to use. Open-source development has shown that individual costs can be lowered if separate parties collaborate on shared artifacts. This form of egalitarian, meritocratic, and self-organizing collaboration, called open collaboration (Riehle et al., 2009), naturally extends to open data. Similar to the open-source workflow, being able to share intermediate artifacts between projects would allow distributed communities of episodic volunteer contributors to collectively increase data quality, motivated by their own reuse. Open data users could collaboratively work to improve data for themselves after it has been published. By doing so, they would not have to rely on data publishers that might be slow to improve their data or have no incentives to provide data in a well-structured format.

In contrast to open-source software development, large-scale open collaboration seems to be uncommon in data engineering by open data users, with most open data projects being completed by small teams (Choi & Tausczik, 2017). It is unclear why open data users do not collaborate as much as open-source developers.

Virtual collaboration plays a major part during data engineering and participants make extensive use of asynchronous collaboration tools like GitHub, Slack, or Email (Choi & Tausczik, 2017; Zhang et al., 2020). Especially open data can be shared and improved among geographically distributed, virtual communities. However, reusing existing software and workflows that have been developed for software engineering does not optimally support data engineering activities. Tools that support virtual and collaborative work during other phases of the data science workflow have shown promise, for example during feature engineering (Smith et al., 2017) or for the creation of labeled data (Reddi et al., 2021). A recent publication by Smith et al., 2021 shows that open-source software development practices can be used during feature engineering as part of a machine learning pipeline. Understanding how open data users collaborate virtually during data engineering will be essential to create workflows and tools that are better adapted to the challenges they face.

Yet, academic research into open data has mainly focused on data publishers. If data engineering by users is described, it is usually seen as just a phase of a larger data science workflow. Therefore, data engineering, as performed by users of open data, is often mentioned in the literature but not described in depth. To support large-scale open collaboration in data engineering across multiple projects, it is necessary to know the participants, their workflows, and the challenges they encounter in their individual projects. We asked the following research question to create an overview of the involved elements:

Research Question: *Which elements of collaboration systems for data engineering by open data users exist, and what are potential challenges?*

We contribute an overview of existing practices, participants, the tools they use, and artifacts open data users create in the course of data engineering collaboration. To do so, we conducted an exploratory literature review to identify the state-of-the-art workflows and processes in projects built on open data. Going beyond the identification of the current reported practices, we elicited the potential challenges to collaboration in open data engineering. Our contributions can be used as a basis for future research into workflow methods and improvements of supporting tools for open collaboration during data engineering.

This paper is structured as follows: First, we review related work in section 2. The research approach for the survey is presented in section 3. Results of the survey are summarized in section 4, followed by a description of their implications beyond the immediate findings in section 5. After a discussion of the limitations in section 6, we summarize the results and point out future research opportunities in section 7.

2. Related Work

To the best of our knowledge, there exist no reviews of how open data users collaborate in data engineering. Mainly, insight is gained across the whole data science workflow from surveys or interviews with data science practitioners, often in commercial settings. If open data is mentioned, the focus of publications is mostly on open data publishers and the work they need to do to provide data of adequate quality.

The challenges of data engineering are an active research topic in corporate environments. Terrizzano et al., 2015 describe what they call ”Data Wrangling” at IBM, highlighting various barriers like privacy or technical issues to data usage. Also at IBM, Zhang et al., 2020 investigate collaboration of data science workers in a large company environment. They find data scientists are highly collaborative, work in small teams and with a variety of tools. However, they point out it is unclear if their results are generalizable outside of the specific corporate environment at IBM.

Previous work by Wang et al., 2019, while mainly focused on work practices of data scientists and their impressions of automated AI, includes a review of academic literature on what roles exist in data science teams and tools that are used during data science activities. They find interactive computing software like Jupyter Notebooks to be a widely used tool in companies like IBM and Netflix. At the same time, they also point out the problem of overly complex tools and missing features to include domain experts in data science teams.

In the context of open data, Choi and Tausczik, 2017 used interviews and survey responses to gain insight into collaboration during open data analysis. Their results show most collaboration happens in small, interdisciplinary groups that mainly build tools to make the use of data easier or reports based on new information from data. They identify that open data analysis is a new phenomenon that has yet to develop standardized norms and practices, as well as the lack of a centralized collaboration platform, as reasons for the fact that large-scale open collaboration is uncommon. While some participants used GitHub, Choi and Tausczik discuss that the platform might lack features and call for further research into how a platform could best support open data analysis.

Zuiderwijk et al., 2014 take a wider ecosystem perspective of open government data, including data publishers. Their work includes a systematic literature review to identify key elements that allow for the publication and use of open data across all stakeholders, including publishers. These elements include releasing data on the internet, being able to search for appropriate data, processing it, and finally using the data and providing feedback to publishers. Additionally, they point out the need for elements to integrate different tools and data sources.

3. Methods

We conducted a systematic literature review (SLR) according to Kitchenham (Kitchenham, 2004). During an initial pilot study, we identified the need for a review because most research in open data focuses on the activities of data publishers. If data engineering is discussed, it is only in the context of a larger data science process and often specific to a domain. To identify collaboration practices that are applicable to all open data users, we had to create an overview from the state-of-the-art literature.

3.1. Search Strategy

We defined an initial search strategy and refined it with information from the pilot study. The pilot study itself consisted of an iterative and broader approach

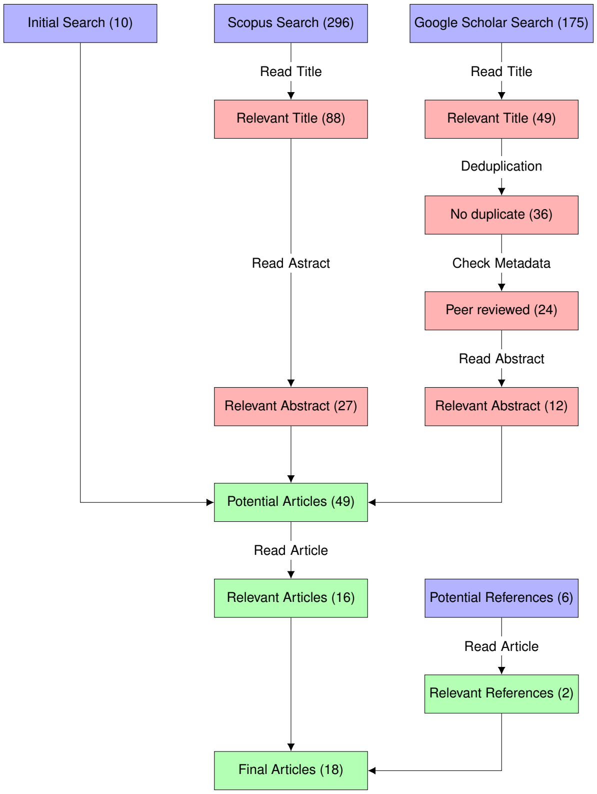

Figure 1. Process of the systematic literature review

to gain familiarity with the literature on collaborative work, data engineering, and open data over a variety of academic search engines. We included articles that were potentially relevant to the research question from these results. Based on the knowledge gained from the pilot study, we defined a systematic search strategy. We retrieved literature from Google Scholar and Scopus to cover a wide range of publications.

An overview of the process is shown in Figure 1. We searched for articles that included *open data* and *workflow*, *process*, *practices* or *participants* or variations thereof.

Scopus offers a comprehensive search interface that allowed us to search in the title and abstract of publications, while still ensuring relevance by making sure the keywords were not too far apart. The keywords in the search string used were:

```
("open data" OR "open-data")
W/5 ("workflow" OR "workflows"
OR "process" OR "processes"
OR "practices" OR "participants")
```

In addition, we limited the results to articles written in either English or German, the type of article only to journals or conference proceedings, and the publication stage to final.

The Google Scholar search was executed similarly. Because Google Scholar does not offer the ability to limit the distance of keywords in abstracts, we restricted the search to paper titles. The search string used was:

```
allintitle:workflow OR workflows OR
process OR processes OR practices
OR participants "open data"
```

For all searches, we only included articles published after 2008 because most publications on open data were created after that time (Purwanto et al., 2020).

We defined explicit inclusion and exclusion criteria:

- **Include** articles that describe data engineering workflows or processes with open data
- **Include** articles reporting on data engineering during a concrete project with open data
- **Exclude** articles that are not peer-reviewed journal or conference papers
- **Exclude** articles exclusively on data publishers
- **Exclude** articles that could not be retrieved in full

Every result of our search algorithm was checked for relevance first by its title, then by its abstract, and finally by skimming the article's full text and applying the inclusion and exclusion criteria. We removed duplicates and any articles that could not be accessed.

During the reading of potential articles, we noted down references that were potentially relevant because they were mentioned in the context of data engineering by open data users. We included these references in the pool of potential articles and verified their relevance by applying the same inclusion and exclusion criteria as for other articles.

3.2. Data Extraction & Synthesis

Our data extraction strategy was aimed at listing all elements of collaboration systems. We set up four shared documents in a sheet management software to track any mention of an activity, participant, tool used, or artifact created during data engineering with open data. We worked iteratively: When an element was mentioned in the current article, we added it to the corresponding list and saved a reference to the source.

When presenting work that builds on open data, most authors do not focus on the exact data engineering activities performed but on the final results. To capture the whole scope of data engineering by open data users, elements were included liberally if they were mentioned in an article, even if they were not part of the main contribution. After every article, we merged entries that had already been identified in previous articles and noted the number of new elements found.

We followed the descriptive data synthesis approach described in Kitchenham, 2004 to provide a broad overview of data engineering in open data, prioritizing including edge cases over a compact summary. We therefore explicitly kept any elements that were only mentioned in few articles but provided new insights to cover the whole breadth of the process.

We grouped elements only when including individual elements did not offer additional insight, mostly for mentioned tools. Here, we merged entries that mentioned different concrete tools of a common type without a clear distinction, but kept any concrete tools that were mentioned specifically. For example, we grouped various mentions of PHP, Python, Java, etc. in the context of implementing software tools into one *general purpose programming languages* entry but kept *Open Refine* as a specific tool because it was mentioned multiple times explicitly.

Descriptions and examples were added for the extracted elements to clarify their meaning, the full descriptions are part of the raw data[1].

In addition, it became apparent during data extraction that a large number of different activities are performed by open data users during data engineering. We considered it important to preserve the detailed separation because the data engineering process is seldom described in detail, especially in the context of open data. However, with the large number of activities identified, we felt it would be valuable to create groups for a clearer overview. We created these groups one by one after all activities had been extracted by considering the list of activities, their descriptions, and examples. Once we felt that every activity was assigned to a matching group, we stopped the addition of new groups.

In a final step, the SLR results were shared with an open data expert working in the domain of open transport data for a member check (see Table 1). Their feedback pointed out some additional activities and distinctions between artifacts but was overall positive and confirmed that they felt the data was complete.

3.3. Concluding the search

Because the goal of this study was to identify the diversity of elements in collaboration systems for open data engineering, we used theoretical saturation as the stopping criterion for the search. Theoretical saturation is considered to be reached when no new insights are gained by analyzing additional data (Bowen, 2008). During our iterative approach to data extraction, we counted the number of new elements added with every article. We considered the data adequate when we did not add any new elements for multiple additional articles.

3.4. Quality Assurance

During writing, we regularly held peer debriefing sessions (Spall, 1998) to ensure the credibility of the results. We discussed qualitatively with two other researchers that were not working on the same research but had experience with the methods we used. Two review sessions were conducted. First, we presented the search strategy of the systematic literature review and the resulting articles, as well as the extracted data. Based on the feedback, we added additional detail to the methods description and discussion of results. In a final peer debriefing session, we focused on the challenges that were identified from the data and how to best present them.

After obtaining the results, the results were discussed with an expert that has practical experience working on multiple open data projects as a form of member checking (Guba, 1981). For this, we created a handout document describing the research goals and methods and asked if we either had identified any elements that should not be included or missed any elements that were part of their practical experience. Based on the comments from the open data expert we then revised the results slightly and explicitly asked if the data seemed complete to which the open data expert confirmed that they had no further comments.

	Method	Participants	Topic
#1	Peer Debriefing	2 Researchers	Search Strategy & Results
#2	Member Check	1 Open data expert	Results
#3	Peer Debriefing	2 Researchers	Identified challenges

Table 1. Feedback methods used

Table 1 shows an overview of feedback sessions, participants, and main topics.

4. Results

We first discuss the search results of the systematic literature review. In addition, the identified elements of open collaboration systems during data engineering by open data users are presented by their categories of participants, activities, tools, and artifacts. As described

[1] Available on Zenodo at https://doi.org/10.5281/zenodo.6598447

in section 3, the list of activities also includes themes that were created by grouping related activities.

4.1. Search results

The search returned 296 results from Scopus and 175 from Google Scholar, as shown in Figure 1. We initially excluded articles by their title, leaving 88 results from Scopus and 49 from Google Scholar.

For results from Google Scholar, we removed duplicates and articles that were not peer-reviewed journal papers or conference proceedings, leading to the exclusion of 25 articles. We then read the abstracts of all remaining articles and kept any that sounded relevant to the research question. After this step, 27 results from Scopus as well as 12 results from Google Scholar were included. Together with the initial search, 49 potentially relevant articles were identified.

The remaining results were read in full and the inclusion/exclusion criteria were applied. During this step, 33 additional articles were excluded, largely because they did not cover data engineering but only later phases of the data science workflow.

In the process of reviewing articles, six potentially relevant references were noted down for future revision. The same inclusion- and exclusion criteria as for other articles were applied, excluding two as not peer-reviewed and two as not relevant. The remaining articles (Lnenicka & Komarkova, 2019; Magalhaes et al., 2013) were included in the final literature pool, but did not contribute newly identified elements.

In the end, we identified a selection of 18 relevant articles.

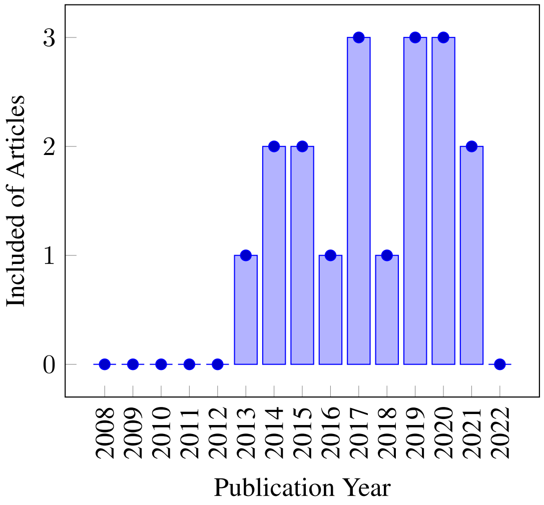

Figure 2. Publication date of included articles

We searched for publications starting in 2008, included articles were published between 2013 and 2021 (see Figure 2) with a slight increase since 2017. Most publications from 2022 could not be included because the original searches were performed during March and April 2022.

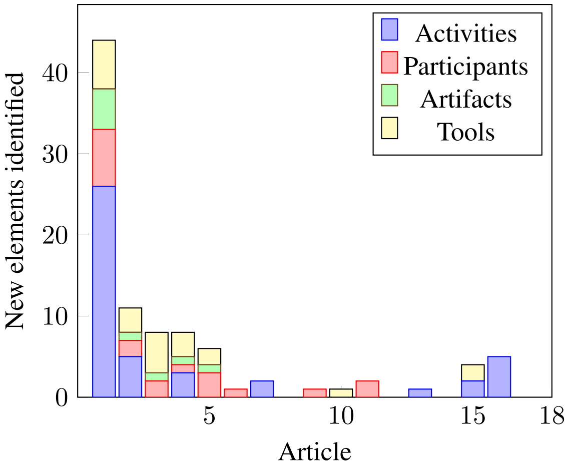

Figure 3. New elements identified by article

We tracked the number of new elements we added with every article (see Figure 3). Because the vast majority of elements were identified in the first articles and later articles contributed no new insights, we considered theoretical saturation to be reached and concluded the review.

The complete search results as well as data extraction are part of the raw data [1]. The raw data also includes sources for all identified elements of collaboration systems presented that have been omitted in the tables for readability.

4.2. Participants

Participants	
Businesses	Mediators
Citizen Scientists	NGOs
Civil Servants	Open Data Experts
Data Scientists	Organisations
Domain Experts	Private Citizens
Goverment Agencies	Researchers
Hackathon Participants	Software Developers
Infomediaries	Startups/Entrepreneurs
Journalists	Students
Legal Advisors	

Table 2. Participants in data engineering, by user role

A diverse group of users participates in collaboration systems for data engineering with open data, as shown in (Table 2). Open governmental data is a focus of academic research, so the list of participants is more detailed for the domain of public administration. It is noteworthy that government agencies and civil servants also act as users of open data, not only as publishers.

Open data allows interested parties insights into political processes, resulting in private citizens, NGOs, and journalists being common participants in open data engineering. These stakeholders are mainly involved in creating reports from data or tools for further insight.

We also identified commercial participants that use open data to build or enhance their products. These include large businesses like IBM that use open data as part of their larger data management strategy, but also startups that may build products based solely on open data. The literature also included references to intermediate entities, called Infomediaries, that offer services based on open data to end users, e.g., companies offering improved open data as a service, either by bundling it or providing processed data.

Open data users come from different backgrounds and view data from different perspectives. On one hand, the use of data creates a number of challenges in itself, requiring input from legal advisors and experts in open data or data science itself. On the other hand, working with open data is a technical challenge, which means software developers are often part of open data projects.

Depending on the context, understanding open data can be complicated and domain experts must be part of the data engineering process. This is especially true for the use of open data by researchers and students when the data might be part of a larger academic project, but also for citizen scientists that want to make sense of a complex problem.

In contrast to general data engineering, open data is often used by hobbyists or amateurs in the context of hackathons, private citizens, or students in university projects. Common to all these users is the low amount of organization and direction, an environment that should be ideal for open collaboration approaches.

4.3. Activities

A wide range of activities is potentially performed as part of data engineering by open data users. Table 3 shows an overview with the larger themes that emerged from the data. Not every activity will be executed during a given project, often only a small subset of activities is needed to make raw data available to use in an application.

In any case, users will need to perform activities related to *acquiring* and *assessing* open data to use. Most often, acquiring data takes the form of searching or discovering data and extracting it by downloading a data set. More complex projects might need to build infrastructure to automatically access data repeatedly. Not all data is easy to extract either, some data publishers require the creation of accounts or impose limits on how often data can be downloaded. After the data is acquired, it must be assessed for appropriate scope and legal compliance. To do so, users often visualize or preview part of the data. From expert feedback, we learned that availability is often a concern for open data users, when relying on an open data source it is important to verify that it will be consistently reachable. This process is necessarily iterative with backtracking whenever a data source lacks the content, license, or availability needed to be useful.

Once appropriate data has been acquired, it can be *improved* or *extended* with additional data. These steps contain a large number of technical activities like changing data format or structure, normalizing values, as well as finding and fixing errors. Additionally, users link data with other data sets and add metadata like data quality indicators. If the data is in a different language it might be required to translate it, either by employing automated translation tools or by hand. Activities that are preparing data for later stages in an ML workflow like feature creation and labeling of data could be considered project-specific and therefore not relevant to general data engineering. They are included here because well-structured, public data sets can be a useful basis for multiple ML projects that have no direct relation (Reddi et al., 2021).

For open data projects that are planned to exist long-term, *maintaining* data becomes a concern. Users need to write documentation about the process to make sure it can be repeated and data can be refreshed if it changes, like e.g. transportation schedules. Archival of open data might be necessary, especially if the underlying data source is unreliable or old data is replaced with new data.

To perform any of these activities it is essential to *understand* the data. This can be a purely technical challenge to learn the data format and structure of the data. Users analyze parts of the data or create small, ad-hoc experiments to gain insights into the data. Often, understanding data also requires understanding the underlying problem domain. Depending on the complexity of the context, this can mean having to ask (and find) domain experts or having to build up domain knowledge.

During all of these activities, open data users *communicate* with different participants in the

Acquire	Assess	Communicate	Extend	Improve	Maintain	Understand
Build Infrastructure	Ensure Anonymity	Ask Publisher	Add Metadata	Aggregate	Archive	Analyze
Discover	Evaluate	Discuss	Create Features	Clean	Document	Ask Experts
Extract	Preview	Find Community	Label	Combine	Refresh	Experiment
Read Documentation	Measure Availability	Find Skilled Users	Rate	Curate		Learn Domain Knowledge
Search	Verify License	Give Feedback	Translate	Enrich		Learn Structure
Select	Visualize / Plot Data	Request Data		Link		
Store		Share Data (Publisher)		Normalize		
Validate		Share Data (Stakeholders)		Reformat		
		Share Information		Repair		
				Structure		

Table 3. Activities performed during data engineering by open data users

ecosystem. They ask questions and provide feedback to data publishers, search for skilled users or domain experts in a community surrounding the data, and share their data and additional information with others. It is noteworthy that interactions with data publishers are expected and open data portals provide avenues to contact them. On the other hand, communicating with the larger community of users that are interested in the same data is less common during data engineering. Regarding other users, activities related to identifying experts and finding other community members are mentioned more often.

4.4. Tools and Artifacts

Tools used	
Auth Providers	Kaggle
Big Data Processing Tools	Notebooks
Blogs / Websites	Official Discussion Board
Command Line Tools	Open Data Repositories
Data Science Libraries	Open Refine
Databases	Sheet Software
Domain Specific Languages	Statistical Computing Languages
Domain Specific Software	Translation Software
General Purpose Languages	Travis
git	Visualization Tools
GitHub	Wikis

Table 4. Tools used during data engineering by open data users

Table 4 shows mentions of tools in literature. We could not identify a standard tool outside of Open Refine which was mentioned multiple times. Depending on the technical skills of project members, employed tools can be self-developed (e.g., based on general-purpose languages) or pre-made applications like Wikis or Sheet Software. After expert feedback, we added custom *Software Applications* as an explicit artifact. We previously assumed open data practitioners collaborate on building their own software in an open-source development process (so the artifact would be *Source Code*) but some applications (e.g., data validation tools) are also developed internally and only shared as closed-source programs.

Visualization tools play an important role in the data engineering workflow because they allow users to quickly evaluate data for quality and scope. When performing the more technical activities for acquiring and improving data, practitioners rely largely on general-purpose programming languages like Python or Java and the surrounding ecosystem of tools like Jupyter Notebooks and GitHub.

Open data repositories are mentioned often, but mainly just as a source of raw data. In contrast to open-source development, where collaboration increasingly happens on GitHub as a central project repository, many different open data repositories exist and data is spread between them. To find experience reports about data, documentation, and feedback, users must visit multiple, disconnected locations like publisher websites, practitioner blogs, or discussion boards.

Created Artifacts	
CI Definitions	Notebooks
Comments on Data	Processed Data
Data Quality Ratings	Raw Data
Documentation	Software Applications
Feedback-/Experience Reports	Source Code
Metadata	

Table 5. Created artifacts by open data users during data engineering

Similarly, open data practitioners do not collaborate on one well-defined, shared artifact. Various artifacts are created as part of data engineering activities (see Table 5) but they mostly are related to metadata or tools to deal with data.

5. Challenges

Open data has the potential for productive open collaboration because the data itself as well as any products resulting from it can be shared freely. Despite this, data engineering in open data is largely considered an activity for data publishers that concludes when the data is made public. We identified a number of potential challenges from the results of the systematic literature review:

- Need for specialized skills but high barriers to participation
- Finding and connecting with other community members
- No standard tools or artifacts
- No well-understood collaboration practices

First, the use of open data requires additional skills, making it more difficult for domain experts to participate. Experience with software development is a vital part of data engineering and software developers are common participants in open data projects. In addition, general-purpose programming languages, as well as statistical computing languages, were among the most mentioned tools for data engineering in our literature review. Aside from software engineering, the required data management skills can impose a barrier as well. Common formats to describe structured open data include semantic web formats like RDF, yet in a recent survey of researchers in Kjærgaard et al., 2020 only 7% of respondents said they were comfortable using it. The challenge will be to lower technical barriers to participation while at the same time staying flexible enough to work with a variety of data sources, formats, and qualities.

A second challenge is connecting members of open data communities. Our results show that the process of understanding data involves learning domain knowledge by finding domain experts or help from a community. Additionally, skilled users have to be identified to help with various barriers from software development to legal advice. Here, the challenge to open collaboration is that users must be able to identify and contact other participants that have a required skill set and are interested in the same data. Due to the public nature of open data, communities can naturally form around an area of interest but it is hard to find other members (Ruijer & Meijer, 2020). Other domains of open collaboration like open-source software development or wikis provide a central location (e.g., GitHub or Wikipedia) for community members to interact. In open data, a similar role could be accomplished by open data repositories, yet they are focused on providing data and less on building communities around common interests.

The lack of a standard artifact is an additional challenge for open collaboration. In open-source software development, community members collaborate on clearly defined artifacts, expressed in source code, like frameworks or libraries. These artifacts have in common that they are not competitively differentiating but instead get used to build additional, potentially closed-source products on top of. By collaborating on an underlying artifact, open-source developers can lower individual costs and with increased generality and quality of the artifact improve their individual applications. We could not identify a similar artifact in open data engineering. Most artifacts described in Table 5 are metadata surrounding the use of open data but not the data engineering steps themselves. It will be a challenge for open data engineering to find an intermediate artifact that is generic enough to be of use for many projects but can be developed collaboratively. Even though the processed open data is an obvious artifact that can be re-shared with the community, it is too static. As identified by Terrizzano et al., 2015, data must be regularly refreshed to be up-to-date, the same is true for regularly released data sets like open transport schedules. An ideal intermediate artifact would be able to cope with changing or newly released data.

Without a shared artifact to collaborate on, data engineering for open data faces the challenge of a fragmented tooling landscape. Currently, traditional software development tools like programming languages and GitHub are common in data engineering (see Table 4). As pointed out by Choi and Tausczik, 2017, these tools lack features that support collaboration on data specifically. Especially during evaluation, participants also use a variety of visualization tools and sheet software to preview the scope and quality of open data. For large-scale open collaboration, a centralized location to collaborate on an artifact will be important. In the open-source approach to software development, this role has been increasingly filled by GitHub for source code and module repositories like npm. Because a more data-focused, well-built project forge has yet to be created, GitHub is currently also used in many open data projects but is missing solutions for e.g., previewing and visualizing data.

As a result of the previous challenges, limited collaboration practices have been developed and adopted during data engineering on open data. Instead, a mindset of data publishers on one side and data users on the other side is common. Participants acquire data as-is and improve it for their own use-case but seldom share the resulting data with the community. If data is released with errors or in inconvenient formats, users provide feedback and quality ratings to publishers but do not work together to improve the data for everyone. One exception are so-called *Infomediaries* (see Table 2) that offer additional services on top of existing data. This mindset difference is in contrast to the open-source approach to software development, where developers are collaborating on shared source code that then is used in their individual projects. While some data engineering activities like evaluating the scope of data might be project-specific, others like finding and fixing errors could be shared by interested parties. A final challenge to open data collaboration will be to identify which activities can be performed by aligned participants and develop collaboration workflows and practices for them.

6. Limitations

Our search was limited to Google Scholar/Scopus as well as by the language of articles, potentially missing out on relevant articles. We noted and reviewed relevant forward references from the original search results to increase our confidence in the completeness of the results. However, the articles identified in the literature search only include academic work while some papers also cited not peer-reviewed content. An additional search of practitioner literature would improve the depth of the review.

We performed descriptive data synthesis for the results of the systematic literature review. Without extracting quantitative data, we can not make statistical inferences about how common or important the identified elements of open collaboration systems in data engineering by open data users are. Given the research goal of identifying the diversity of elements, this was appropriate and allowed us to contribute a descriptive overview. At later points in time, with an extended search, other forms of qualitative data analysis could also be used. Ideally, quantitative data should be surveyed from open data practitioners instead.

A threat to validity in the form of bias could exist because large parts of the academic literature on open data is related to open government data. Because our goal is to identify as many elements as possible, and we are not attempting quantitative data synthesis, the threat is mitigated. However, the potential to miss elements from other domains exists. We used expert feedback from an open transport data practitioner Table 1 to increase our confidence that we captured the whole breadth of the data engineering process.

7. Conclusion

In summary, we set out to identify elements of collaboration systems for data engineering by open data users and point out potential challenges.

We have performed a systematic literature review and descriptive data synthesis to find elements of open collaboration systems in data engineering. Our results show that open data users come from many domains, with varying technical skills, and perform a large number of activities. We could find different tools and artifacts, but no standard practice of collaboration across open data engineering.

We identified a number of potential challenges to open collaboration in data engineering: High barriers to participation but the need for specialized skills, identifying and connecting with a larger community, the lack of standard tooling and artifacts as well as missing collaboration practices.

These challenges are especially relevant for large-scale, virtual collaboration that has the potential to be very effective in the context of open data projects. Working virtually with unknown members in a larger community exacerbates the identified challenges and makes standard tooling and practices even more important. Our results will be the basis for the development of an open collaboration workflow method and supporting tool that allows data engineers to collaborate in a geographically dispersed and asynchronous manner.

Additionally, we plan to extend our description of open collaboration systems in data engineering and verify the discussed challenges. To do so, we will conduct interviews with open data practitioners as well as industry partners in further work.

Acknowledgments

The authors would like to thank the open data expert for their feedback, both in informal talks as well as in reviewing the paper results. Additionally, the thoughtful comments from the anonymous reviewers helped us improve the paper.

References

Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative research*, *8*(1), 137–152.

Choi, J., & Tausczik, Y. (2017). Characteristics of collaboration in the emerging practice of open data analysis. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ*, *29*(2), 75. https://doi.org/10.1007/BF02766777

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1–26.

Kjærgaard, M. B., Ardakanian, O., Carlucci, S., Dong, B., Firth, S. K., Gao, N., Huebner, G. M., Mahdavi, A., Rahaman, M. S., Salim, F. D., Sangogboye, F. C., Schwee, J. H., Wolosiuk, D., & Zhu, Y. (2020). Current practices and infrastructure for open data based research on occupant-centric design and operation of buildings. *Building and environment*, *177*(106848), 106848. https://doi.org/10.1016/j.buildenv.2020.106848

Lnenicka, M., & Komarkova, J. (2019). Big and open linked data analytics ecosystem: Theoretical background and essential elements. *Government information quarterly*, *36*(1), 129–144. https://doi.org/10.1016/j.giq.2018.11.004

Magalhaes, G., Roseira, C., & Strover, S. (2013). Open government data intermediaries: A terminology framework. *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, 330–333. https://doi.org/10.1145/2591888.2591947

Purwanto, A., Zuiderwijk, A., & Janssen, M. (2020). Citizen engagement with open government data. *International journal of electronic government research*, *16*(3), 1–25. https://doi.org/10.4018/ijegr.2020070101

Reddi, V. J., Diamos, G., Warden, P., Mattson, P., & Kanter, D. (2021). Data engineering for everyone. *CoRR*, *abs/2102.11447*. https://arxiv.org/abs/2102.11447

Riehle, D., Ellenberger, J., Menahem, T., Mikhailovski, B., Natchetoi, Y., Naveh, B., & Odenwald, T. (2009). Open collaboration within corporations using software forges. *IEEE Software*, *26*(2), 52–58. https://doi.org/10.1109/MS.2009.44

Ruijer, E., & Meijer, A. (2020). Open government data as an innovation process: Lessons from a living lab experiment. *Public performance & management review*, *43*(3), 613–635. https://doi.org/10.1080/15309576.2019.1568884

Smith, M. J., Cito, J., Lu, K., & Veeramachaneni, K. (2021). Enabling collaborative data science development with the ballet framework. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2), 1–39. https://doi.org/10.1145/3479575

Smith, M. J., Wedge, R., & Veeramachaneni, K. (2017). FeatureHub: Towards collaborative data science. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 590–600. https://doi.org/10.1109/DSAA.2017.66

Spall, S. (1998). Peer debriefing in qualitative research: Emerging operational models. *Qual. Inq.*, *4*(2), 280–292.

Terrizzano, I. G., Schwarz, P. M., Roth, M., & Colino, J. E. (2015). Data wrangling: The challenging yourney from the wild to the lake. *CIDR*.

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359313

Zhang, A. X., Muller, M., & Wang, D. (2020). How do data science workers collaborate? roles, workflows, and tools. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW1), 1–23. https://doi.org/10.1145/3392826

Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information polity*.