Challenges to Open Collaborative Data Engineering

Philip Heltweg, Dirk Riehle
Friedrich-Alexander-Universität Erlangen-Nürnberg

Minitrack on Virtual Collaboration, Organizations, and Networks
@ HICSS-56

Authors

Philip Heltweg

Friedrich-Alexander-Universität Erlangen-Nürnberg

philip@heltweg.org

https://heltweg.org

Prof. Dr. Dirk Riehle, M.B.A.

Friedrich-Alexander-Universität Erlangen-Nürnberg

dirk.riehle@fau.de

https://dirkriehle.com

Motivation

- “Open data and content can be **freely used, modified, and shared** by anyone for any purpose.”, The Open Definition
- Poor data quality, technical challenges for data users
- Data users could collaboratively improve data
  - Similar to open-source software
  - Potential for large-scale, virtual collaboration
- Literature focuses on data publishers
  - Unclear how open data users do data engineering and what challenges they face

Research question & Contributions

“Which elements of collaboration systems for data engineering by open data users exist, and what are potential challenges?”

Contributions

- Overview of activities, participants, tools and artifacts in data engineering on open data
- Challenges to open collaboration during data engineering on open data

Methods Overview

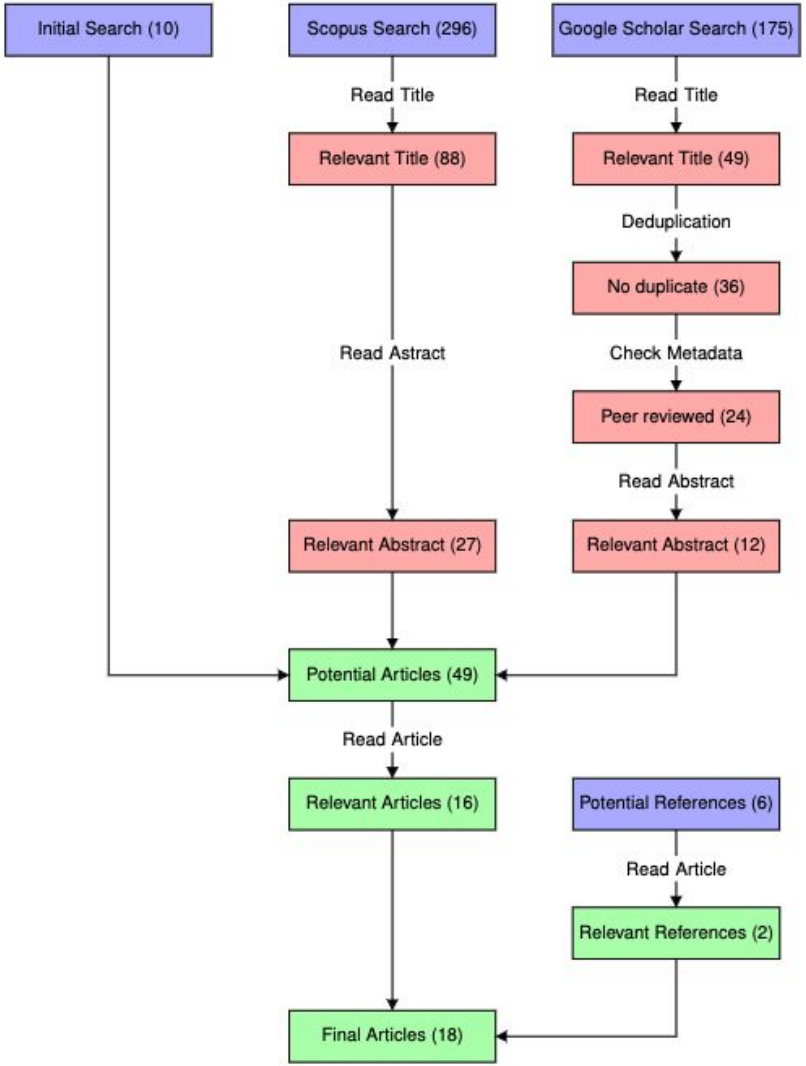
- Structured literature analysis according to Kitchenham, 2004
  - Search until theoretical saturation (Bowen, 2008)
  - Descriptive data synthesis
- Quality Assurance
  - Peer Debriefing according to Spall, 1998
  - Member check with open data practitioner (Guba, 1981)

Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1–26.
Spall, S. (1998). Peer Debriefing in Qualitative Research: Emerging Operational Models. Qualitative Inquiry: QI, 4(2), 280–292.
Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: a research note. Qualitative Research: QR, 8(1), 137–152.
Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. ECTJ, 29(2), 75.

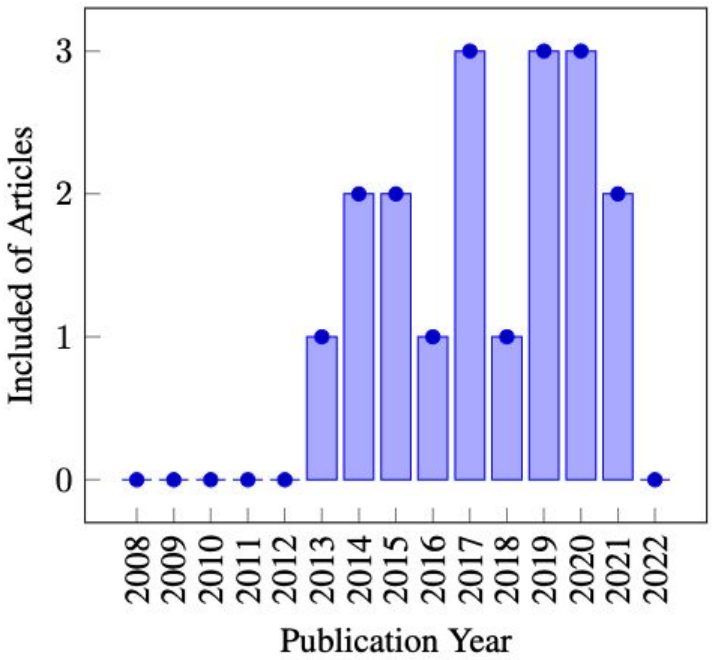
SLR Search Strategy

- Sources: Google Scholar, Scopus, Forward references
- Search terms: “*open data*”, “*workflow*”, “*process*”, “*practices*”, “*participants*” (and variations)
- Include
  - Articles describing projects including data engineering with open data
  - After 2008 (after Obama’s Open Government Initiative (Purwanto et al., 2020))
- Exclude
  - Not peer-reviewed
  - Exclusively on data publishers
  - No access

Purwanto, A., Zuiderwijk, A., & Janssen, M. (2020). Citizen engagement with open government data. International Journal of Electronic Government Research, 16(3), 1–25.


Initial Search (10)
Scopus Search (296)
Google Scholar Search (175)
Read Title
Read Title
Relevant Title (88)
Relevant Title (49)
Deduplication
No duplicate (36)
Read Astract
Check Metadata
Peer reviewed (24)
Read Abstract
Relevant Abstract (27)
Relevant Abstract (12)
Potential Articles (49)
Read Article
Relevant Articles (16)
Potential References (6)
Read Article
Relevant References (2)
Final Articles (18)

Search Results



- 487 articles considered
- 18 relevant articles between 2013 - 2021 (search executed in April 2022)
- Raw search result data available from Zenodo (https://doi.org/10.5281/zenodo.6598447)

Activities

- **Seven categories** with 44 activities
  - Acquire Data
    - Search, Extract, Store…
  - Assess Data
    - Evaluate, Visualize, Verify license…
  - Communicate about Data
    - Find skilled users, Give feedback, Request data…
  - Extend Data
    - Add metadata, Rate, Translate…
  - Improve Data
    - Clean, Normalize, Combine…
  - Maintain Infrastructure
    - Archive, Document, Refresh…
  - Understand Data
    - Ask experts, Experiment, Learn domain knowledge…

Participants

- Often: Open government (Government agencies, Journalists, NGOs…)
- Some: Commercial (Startups, Large businesses like IBM…)
- Open data often used by non-professionals
- Participants with diverse backgrounds and skillsets

Participants lack required **domain knowledge** and **technical expertise** for data engineering

Tools & Artifacts

- Tools depend on technical skill
  - self-developed to pre-made
- Popular: Open Refine, Open data repositories (e.g., CKAN), visualization tools for data exploration
- Created artifacts mainly metadata, documentation or software
  - seldom improved data itself

No standard tools or artifacts

Challenges to open collaborative data engineering (1/2)

- **Need for specialized skills but high barriers to participation**
  - Technical barriers limit involvement by domain experts
  - For example, RDF is a common data format, yet only few researchers are comfortable using it (Kjærgaard et al., 2020)
- **Finding and connecting with other community members**
  - Communities can naturally form around open data but it is hard to find other members (Ruijer & Meijer, 2020)

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.
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.

Challenges to open collaborative data engineering (2/2)

- **No standard tools or artifacts**
- **No well-understood collaboration practices**
  - Tools and practices from software engineering are reused but lack important features specific to collaboration on data (Choi & Tausczik, 2017)

Choi, J., & Tausczik, Y. (2017, February). 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. CSCW '17: Computer Supported Cooperative Work and Social Computing, Portland Oregon USA.

Outlook

- **Extend identified challenges** (survey and interviews with open data practitioners)
- Suggest and evaluate standard collaboration **practices** for data engineering
- Implement **tools** to support collaborative data engineering