Tools and Ecosystems for Open Control Systems Data at ESS

Runeson, Per; Söderberg, Emma

2021

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Runeson, P., & Söderberg, E. (2021). *Tools and Ecosystems for Open Control Systems Data at ESS*. (Technical report; No. 106). Lunds Universitet/Lunds Tekniska Högskola.

Total number of authors:
2

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Tools and Ecosystems for Open Control Systems Data at ESS

Per Runeson and Emma Söderberg

January 25, 2021

Summary

We see a potential for ESS to benefit from sharing alarm data for machine learning applications, by establishing data ecosystem. To achieve these goal, ESS must make a strategic decision to pursue data sharing, establish a technical platform to share a digital twin, continue advancing the data ecosystem, and work towards a standardized reference model.

1 Introduction

This is the summary report from the work package 1 (WP1) ESS Control System Data Lab (ESS-CSDL) pre-study. It contains preliminary recommendations regarding *data*, *tools*, ecosystem *governance* and a *vision* for the upcoming steps of ESS-CSDL.

The recommendations are based on our preliminary concept definitions [8], exploration of factors relevant for data ecosystems in focus groups [10] (both from preceding projects), the validation of these factors against the ESS-CSDL and two other emerging data ecosystems, Road Data Lab (RoDL) from the automotive domain and JobTech from the job market domain [9], an exploratory study of small scale data engineering for machine learning [2], and literature on data ecosystems, particularly through the lens of a mapping study by Oliveira et al. [4].

2 Data ecosystems

Data ecosystems is an emerging concept in its infancy [4]. In their mapping of existing research, Oliveira et al. found 29 papers spread over a wide range of research topics, indicating a scattered and immature field. They also conclude that the definition of the term is not yet stable. Thus, we choose to adapt a definition from software ecosystems [3] and open government data [11], in which each of the bold-faced terms is elaborated below:

> *A **data ecosystem** is a networked community of **actors** (organizations and individuals), which base their relations to each other on **a common interest**, supported by an underpinning **technological platform** providing **data shared within the ecosystem**. Actors collaborate on the data and **boundary resources** (e.g., software and standards), through the exchange of information, resources and artifacts.*

In our study of three emerging **data ecosystems**, we observed them to be either [9, Section 5.2.4]: 1) agency-driven, 2) business-driven, or 3) community-driven. ESS-CSDL is – despite its public funding – aiming to be a commercially viable ecosystem *(business-driven)*, which provides business value for its stakeholders, although it in the initiation phase needs support from public innovation agencies.

The aim of the Control System Machine Learning (CSML) program at ESS is defined in the project application "to drive development within artificial intelligence and machine learning for society through large scale research facilities". The current **actors** or stakeholders are of three kinds: 1) DESY as a similar research facility, 2) companies providing automation equipment, and 3) companies in the process industry. The **common interest** around which the ecosystem is founded is to promote this aim. The primary means is to share experience and **data in the ecosystem**. Note that the exchange of experience is considered as important by the actors as the exchange of data, at least in this initial stage of the ecosystem. Thus, it is important to ensure that this is made possible though strategic decisions, and clear in the communication with current and future stakeholders, as further discussed in the data policy report [6].

Recommendation 1: ESS-CSDL must anchor and communicate the strategic decision of sharing data to enable the business benefits of the data ecosystem.

3 Tools for data ecosystems

Data ecosystems need **technological platforms** (or tools) for data collection, enhancement and distribution. There are open source tools of open data portals, for example CKAN[1], which are used for open government data and for open research data. The ecosystem also needs a communication and collaboration platform, sharing issues about the data, making the governance decisions transparent etc.

Albeit ESS being a huge investment, producing enormous volumes of data, ESS-CSDL does not produce large volumes of alarm data in its initial phase. Neither are there any immediate needs for real time data exchange – except for what is already in place[2]. Further, the number of actors in the ecosystem is a handful at the beginning. Thus, with respect to data volumes, existing file sharing and communication tools within ESS suffice for the data sharing needs. However, as the data must be accessible from different organizations, the access rights must be more easily manageable. We therefore propose that ESS-CSDL creates a solution for data storage and access for small to medium scale data to be shared externally under access control, similar to a digital twin model of the alarm system. Figure 1 shows an architectural sketch of such a solution, which implementation is outlined in the data policy report [6].

In the exploratory study of small scale data engineering [2] carried out in this project, it was stressed that the agility in the tools and methods for data engineering were more important for small scale data engineering, than power and capacity typically required when processing large volumes of data. Especially data provenance, for tracing of data versions when building data pipelines, and facilitation of a shared understanding of what data means, was deemed important. In creating a solution for data storage for ESS-CSDL, we propose that this solution should support management of data provenance and meta-data.

Recommendation 2: ESS-CSDL should establish a technical solution for data sharing and communication, to ensure that the access rights may be organized more easily.

[1] `https://ckan.org`

[2] `https://pos.esss.lu.se/`

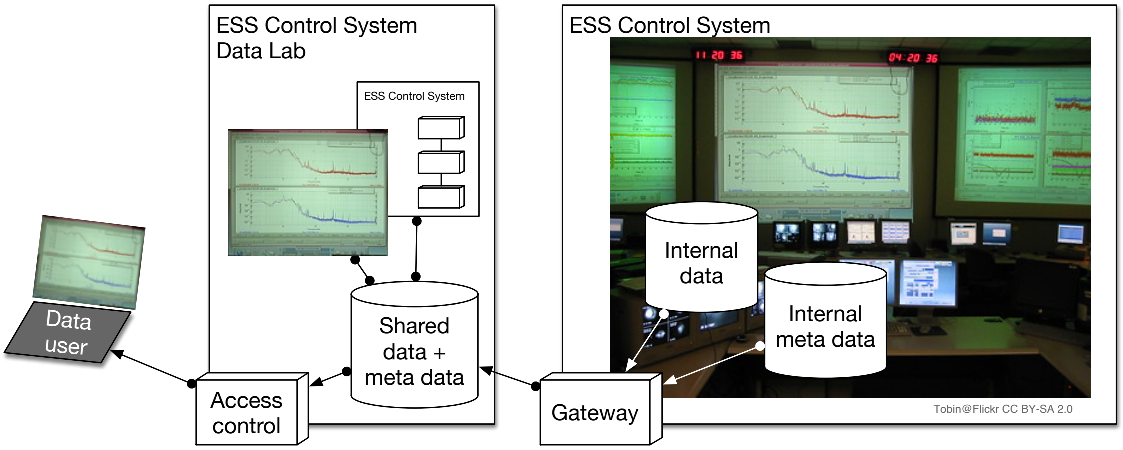

Figure 1: *Architecture of technical platform for data sharing. An external data user has access to a the data lab, which holds data and meta data, mirrored from the ESS-internal control systems.*

4 Ecosystem governance

Governance of a data ecosystem involves the "[e]ntirety of rules, responsibilities and decision-making rights affecting the behavior and interaction of the actors of the platform ecosystem" [5]. This implies setting and upholding the rules for the ecosystem.

We identify three major types of ecosystem governance [9, Section 6]: 1) keystone-centric, 2) consortium-based, and 3) community-based. ESS-CSDL is a *keystone-centric* ecosystem as "actors are organized around a keystone actor, which is directly or indirectly responsible for providing much of the data. ... Hence, the keystone actor should be responsible for monitoring, evaluating, making decisions, and taking actions" [4]. The ESS-CSDL is established based on the alarm data of ESS, and thus ESS *is* the keystone of the ecosystem.

It may evolve towards a consortium-based structure while maturing and including more active members, but currently it is keystone-based. As a consequence, ESS-CSDL must take the leadership to advance the ecosystem further.

Recommendation 3: ESS-CSDL must continue to advance the data ecosystem and ensure there is a common interest among the stakeholders.

5 Vision for the ecosystem

During the work with ESS Alarm Data [1] it became apparent that the metadata and reference models are **boundary resources** which are critical elements for understanding and later sharing data. In other fields, work towards standardization underpin the data ecosystem, for example in public transport [7]. Thus, the work towards establishing generic reference models for Industry 4.0 alarm systems, making them *de facto* or *de jura* standards, is both a contribution from the emerging data ecosystems, and a condition for its survival and growth.

Recommendation 4: ESS-CSDL should continue to work towards standardized reference model for Industry 4.0 alarm systems.

Based on the above recommendation, we envision a control system data ecosystem as outlined in Figure 2, with multiple data providers and users, both suppliers and control system owners.

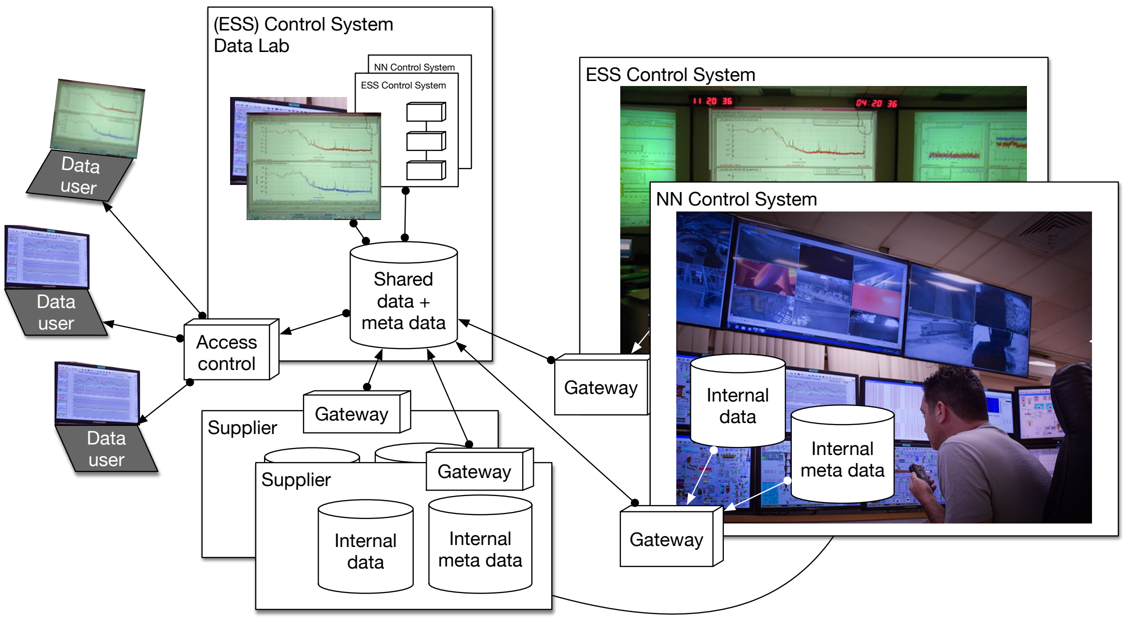

Figure 2: *A future control system data ecosystem with multiple data providers and users.*

References

[1] Per Andersson, Jan Eric Larsson, and Karin Rathsman. ESS control system data lab – work package 2 final report. Technical Report ESS-3216740, ESS, 2021.

[2] Anton Engström. Towards agile data engineering for small scale team. Master thesis LU-CS-EX 2020-40, Lund University, Department of Computer Science, 2020.

[3] Slinger Jansen. A focus area maturity model for software ecosystem governance. *Information and Software Technology*, 118, 2020.

[4] Marcelo Iury S. Oliveira, Glória de Fátima A. Barros Lima, and Bernadette Farias Lóscio. Investigations into data ecosystems: a systematic mapping study. *Knowledge and Information Systems*, 61(2):589–630, 2019.

[5] Boris Otto and Matthias Jarke. Designing a multi-sided data platform: findings from the international data spaces case. *Electronic Markets*, 29(4):561–580, 2019.

[6] Karin Rathsman. ESS control system data licenses: Policy and strategy for the future – work package 3 deliverable. Technical Report ESS-3192568, ESS, 2021.

[7] Daniel Rudmark. Open data standards: Vertical industry standards to unlock digital ecosystems. In *53rd Hawaii International Conference on System Sciences, HICSS 2020, Maui, Hawaii, USA, January 7-10, 2020*, pages 1–10. ScholarSpace, 2020.

[8] Per Runeson. Open collaborative data – using OSS principles to share data in SW engineering. In *41st ACM/IEEE International Conference on Software Engineering, New Ideas and Emerging Results (NIER) track*, pages 25–28, Montreal, Canada, May 2019.

[9] Per Runeson, Johan Linåker, and Thomas Olsson. Open data ecosystems – an empirical investigation into an emerging industry collaboration concept. *Journal of Systems and Software*, Submitted, 2021.

[10] Per Runeson and Thomas Olsson. Challenges and opportunities in open data collaboration – a focus group study. In Antonio Martini and Manuel Wimmer, editors, *Euromicro Conference on Software Engineering and Advanced Applications*, pages 205–212. IEEE, 2020.

[11] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, 19(1, 2):17–33, 2014.