Data Sovereignty for AI Pipelines: Lessons Learned from an Industrial Project at Mondragon Corporation

Marcel Altendeitering
Julia Pampus
marcel.altendeitering@isst.fraunhofer.de
julia.pampus@isst.fraunhofer.de
Fraunhofer ISST
Dortmund, Germany

Felix Larrinaga
Jon Legaristi
flarrinaga@mondragon.edu
jon.legaristi@alumni.mondragon.edu
Mondragon Unibertsitatea
Arrasate-Mondragon, Spain

Falk Howar
falk.howar@tu-dortmund.de
TU Dortmund University
Dortmund, Germany

ABSTRACT

The establishment of collaborative AI pipelines, in which multiple organizations share their data and models, is often complicated by lengthy data governance processes and legal clarifications. Data sovereignty solutions, which ensure data is being used under agreed terms and conditions, are promising to overcome these problems. However, there is limited research on their applicability in AI pipelines. In this study, we extended an existing AI pipeline at Mondragon Corporation, in which sensor data is collected and subsequently forwarded to a data quality service provider with a data sovereignty component. By systematically reflecting and generalizing our experiences during the twelve-month action research project, we formulated ten lessons learned, four benefits, and three barriers to data-sovereign AI pipelines that can inform further research and custom implementations. Our results show that a data sovereignty component can help reduce existing barriers and increase the success of collaborative data science initiatives.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; • **Software and its engineering → Data flow architectures**.

KEYWORDS

data sovereignty, collaborative AI, lessons learned, AI engineering

ACM Reference Format:
Marcel Altendeitering, Julia Pampus, Felix Larrinaga, Jon Legaristi, and Falk Howar. 2022. Data Sovereignty for AI Pipelines: Lessons Learned from an Industrial Project at Mondragon Corporation. In *1st Conference on AI Engineering - Software Engineering for AI (CAIN'22), May 16–24, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3522664.3528593

1 INTRODUCTION

The prevalence of machine learning (ML) and artificial intelligence (AI) solutions is growing exponentially and has moved from hype to reality [6, 27]. Vast amounts of available data have led to a democratization of AI and companies of different sizes and from different industries have started their own AI initiatives and experiments [10]. However, despite their broad adoption, AI and ML have not yet delivered on their promises in industrial practice [27]. The shift from prototypical projects to production-ready development is challenging, and many promising initiatives remain insular and isolated from other processes and systems [10, 27].

Since there is "no AI without data" [27] (p.98), data related challenges, including management, governance, and democratization, are the major obstacles for successful AI initiatives [5, 27]. They hinder organizations from leveraging potentially useful data sets and from fully exploiting the benefits of cooperatively working on joint AI workflows [27]. Hereby, the establishment of collaborative AI (e.g., a car manufacturer and its supplier working on a shared AI model) could help overcome insular initiatives and take full advantage of the benefits offered by AI.

The work on collaborative AI pipelines requires the ability to easily share data with partners [15, 44]. However, current processes of data sharing are often lengthy, cumbersome, and can result in the establishment of 'data governance anti-patterns' [44]. When external partners like consultancies or suppliers are involved, the data processing is particularly difficult and requires a significant amount of organizational, technical, and legal clarification [44]. Even within a single organization, data is often treated as an asset for individual business functions and only shared reluctantly with other departments and projects [27]. Furthermore, there is often no technical guarantee for what purpose data is used once it has been shared with others. This leads to situations in which organizations rather not share their data to avoid losing control [15].

Several technological components that offer data sovereignty functionalities have emerged from science and practice to address these problems (e.g., [17, 21, 23, 34]). However, there is still limited research on their applicability in existing AI pipelines and a concrete socio-technological context. Moreover, the current body of literature lacks the prescriptive knowledge needed for implementing data-sovereign AI pipelines. To understand how such pipelines can be realized, we formulated three research questions:

- RQ1: What practices are applied to realize data-sovereign AI pipelines?
- RQ2: What are the benefits of data sovereignty in AI pipelines?
- RQ3: What are the barriers to data-sovereign AI pipelines?

To investigate the proposed research questions, we conducted a twelve-month empirical action research study on a real-world

CAIN'22, May 16–24, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9275-4/22/05…$15.00
https://doi.org/10.1145/3522664.3528593

AI project at Mondragon, a large Spanish corporation. We closely collaborated with Mondragon on the adoption of a data sovereignty solution in an industrial AI setting. Throughout the project, we qualitatively analyzed and systematically reflected on the design decisions and challenges we experienced, abstracted these, and formulated generalizable findings. For the presentation of our results, we used the recent data challenges model by Gröger [27] as a framework. The main contributions of our study are the presentation of generalized lessons learned and the presentation of benefits and barriers to realizing data-sovereign AI pipelines. This way, our study contributes to answering some of the challenges for engineering AI systems raised by recent studies [6, 10, 27, 40, 44]. Our study, furthermore, paves the way for further research in the area of data-sovereign AI pipelines by highlighting issues that require further research attention. Finally, practitioners can use our findings to inform their own AI architectures and solution designs and to establish data sovereignty in their pipelines.

The remainder of this paper is structured as follows. First, we present the theoretical background in Section 2. In Section 3, we outline the methodological approach we followed for our research and describe our course of actions. We present and discuss our findings in the form of lessons learned, benefits, and barriers in Section 4. In Section 5, we conclude our study by summarizing our contributions, limitations, and outline paths for future work.

2 BACKGROUND & RELATED WORK

High quality data sets are vital for the success of any AI initiative and a highly valuable asset for companies [27, 52]. However, only through enrichment with additional information and the collaborative processing, AI initiatives can leverage their full economic value and lead to innovation [27, 29]. Thus, "data exchange between [and within] companies is an essential feature of digitization and data economy" [29] (p.1) – always facing challenges of data management, quality, or governance [27, 35].

Hereby, the most significant problem is that the data owners are often not comfortable sharing their data. Even if contracts and agreements exist, which state how the shared data will be used, the data provider can never be sure of how and for what purpose the shared data will actually be used [13], and is constantly "afraid of losing control" [15] (p.1). In particular, technically ensuring the interests of the data provider and, consequently, building trust is challenging. With regard to this, we want to focus on *data sovereignty* in the following.

2.1 Data Sovereignty & AI

Since *data sovereignty* is a very young field of research, there are different definitions according to the respective domain and area of focus [30, 66]. Jarke et al. [36] (p.550) describe the term data sovereignty in the context of data ecosystems as follows:

"Data sovereignty refers to the self-determination of individuals and organizations with regard to the use of their data. In contrast to data privacy as defined, e.g., in the European General Data Protection Regulation (GDPR), which sees the citizen in a rather passive role to be protected against powers they cannot confront on an equal footing, data sovereignty aims at enabling «data richness» by clearly negotiated and strictly monitored data usage agreements".

This means that the data owners and providers can define terms and conditions for the exchange and use of their data [52, 58], and that these interests are ensured along the entire data supply chain. Data sovereignty enables the sharing of data between multiple organizations or within a larger data space and should be considered in such scenarios. For example, an automotive supply chain consists of multiple actors who mutually benefit from shared data. A car manufacturer could share data on its material stock with different suppliers to avoid a stop of production. To ensure data sovereignty, the manufacturer could specify that the data is only accessible for a certain amount of time and deleted afterwards.

The literature shows that the need for sovereign data sharing does not solely exist in one domain, but can be found across domains in both industry and research: logistics [24], energy [35], mobility [18, 36], health [15, 36], or smart city [54]. Wherever data assets exist and are gathered to create value, the same questions and problems [26] arise: How can trust be established between data sharing partners? How can data be exchanged interoperably? How can data sovereignty be ensured? The challenge of answering these questions becomes even greater when data should be shared with an unknown number of potentially interested parties and across different legacy systems.

"Industrial IoT systems are characterized by data flowing from sensors to services and applications and possibly back to actuating devices. These data flows span several physical platforms, including resource-constrained sensors, mobile devices, and cloud backends" [59] (p.289). Thereby, technical challenges are already predetermined. Moreover, the use of heterogeneous data processing applications and AI algorithms is the key to heterogeneous and complex data flows that need to be made auditable and transparent according to data sovereignty principles.

Many systems merely focus on the implementation of access control by restricting access to data and e.g., linking it to an authentication mechanism. However, this way, the actual use of the data cannot be viewed and monitored [45, 59]. Following this, there is a need to define and explicitly delineate usage control, which "extends access control by the dimension of time and is able to continuously monitor and control the usage of resources such as files or services" [59] (p.289). With an increasing number of data assets to be transferred between sensors and cloud infrastructure in distributed systems, the implementation and integration of usage control as a technical implementation of data sovereignty is a challenge [59], but nevertheless a necessity [66].

2.2 Data Sovereignty Solutions

To address open questions in the area of data sovereignty and the development of technical solutions, projects such as FIWARE [46], the International Data Spaces (IDS), and Gaia-X have been built up.

The IDS, formerly Industrial Data Space, comprises multiple projects that have been driven since the end of 2014 [52] in cooperation with meanwhile more than 100 companies from industry and research [7], and since 2016 under the organization of the International Data Spaces Association (IDSA) [52]. Its goal is to define and specify the establishment of sovereign data ecosystems in the form of multi-sided platforms [51] and common governance rules, and

to create a suitable environment for appropriate software developments [35]. These ecosystems focus primarily on peer-to-peer data provisioning, exchange, and processing [2].

The IDS Reference Architecture Model [52] describes the ontology [9] and the key components within a data ecosystem that are required to create a "secure, trusted, and semantically interoperable" [9] (p.1) data exchange between two stakeholders. The IDS differs from other projects by focusing on enriching data with terms of usage (hereafter *usage policies*) and the automated negotiation of these.

The core of the IDS data ecosystem is the IDS Connector, which is responsible for the exchange of all necessary information: data assets, meta information, and data processing applications. It serves as a gateway to existing proprietary systems and encapsulates their structure, functionalities, and APIs from external environments [24, 48]. The IDS Connector offers the possibility to define and provide usage policies and data flows, and negotiates agreements within a given framework to achieve an agreed consensus. It also ensures that data is transferred and processed all the way from the data source on the provider side to the data sink on the consumer side in accordance with the negotiated agreement. Therefore, it is able to intercept data processing applications, prohibit access, and, if necessary, delete data. The IDS define their own ontology for this purpose, based on standards such as the Data Catalog Vocabulary (dcat) and the Open Data Rights Language (ODRL) [9].

In addition to the IDS Connector, the IDS define central systems for cataloguing capabilities, a trusted identity management, a transparent monitoring of all information flows, the possibility to extend the systems with use case specific data processing applications, and the option to enrich the used data model.

Besides defining the structure of a data ecosystem, the IDS also specifies the authenticated and encrypted communication protocol that is used by the technical components. In this context, every company should be able to enter an existing data ecosystem with their systems and applications without any major entry barrier. This requires an appropriate identification and the deployment of an IDS Connector with a connection to existing systems. Subsequently, data offers can be created and data from other connectors can be consumed.

Gaia-X is focusing on building a distributed data infrastructure in Europe [11]. In addition to adopting IDS concepts for sovereign data exchange, Gaia-X focuses on the creation of common standards, trusted identity management, and a technology agnostic architecture that focuses primarily on cloud infrastructures and includes federation services [11]. Hence, the goals and core aspects of the European [16] data strategy, which aims at balancing societal, economic, and private interests and power in a fair data ecosystem, is to be realized.

Currently, one of the most advanced reference implementations of an IDS Connector is the Dataspace Connector (DSC) from the Fraunhofer Institute for Software and Systems Engineering (ISST). As open source software, provided by the IDSA organization on GitHub [34], the DSC is recently being used in projects like the Mobility Data Space [18], the Energy Data Space [35], and the Bauhaus.MobilityLab [33].

Driven by the success and adoption of the development of this software, an Eclipse project was launched in summer 2020 by a group of eight organizations. Administered by the Eclipse organization, the Dataspace Connector will be further developed as open source software that aims to provide a highly scalable, modular, and extensible framework for sovereign data exchange [23]. It primarily addresses new challenges such as connection and interoperability with multiple data spaces and identity management across multiple jurisdictions [11]. Core issues, that are also reflected in projects like IDS and GAIA-X, being addressed are data sharing and data sovereignty in cloud-native environments.

The modular, secure system offers fully asynchronous, highly available, and permanently auditable processes. Similar to systems like the Open Policy Agent (OPA) [17], LUCON [59], and MYDATA [31], it strives for cloud-agnostic policy management and enforcement to ensure data sovereignty.

2.3 Related Work

To the best of our knowledge, no study has yet investigated how data sovereignty can be integrated in collaborative AI pipelines. Hence, no systematic explanatory knowledge exists on the implementation of such pipelines. However, apart from AI initiatives, there are several other projects that have implemented data sovereignty principles. For example, Alonso et al. [3] describe the implementation of the IDS Reference Architecture for Industry 4.0 scenarios in the FIWARE European project [21]. In the discussed use case, a pilot was developed to improve the manufacturing process of a factory. It obtains data from a milling machine and a coordinate-measuring machine via an IDS Connector. Afterwards, the data is analyzed in two systems for maintenance and quality control. All three IDS Connectors use a central identity access management system and act in their role as data provider and consumer.

Another IDS Connector reference implementation is the Trusted Connector [60] developed by the Fraunhofer Institute for Applied and Integrated Security (AISEC). To create trust in an IDS ecosystem, multiple requirements must be fulfilled: On the one hand, a trusted identity provider and on the other hand, an IDS Connector implementation with a secure software stack are required. Brost et al. [12] define such a system from the hardware layer, through the kernel and virtualization layers, to the container layer. While primarily focussing on IoT systems, the Trusted Connector implements the specified layers based on the Open Service Gateway Initiative (OSGi) framework to execute all applications as "isolated service or a service bundle in a separate execution environment" [12] (p.44). Data flows can be defined and controlled using Apache Camel [22] to ensure that usage control is implemented at each stage of data processing.

3 THE ACTION RESEARCH CASE STUDY

To investigate the proposed research questions, we conducted an action research case study in collaboration with Mondragon. Action research is a well suited research method for industry-academia collaborations in software engineering and helps "to make academic research relevant" [8] (p.94). It can assist to develop innovative solutions and gain an in-depth understanding of novel phenomenons in real situations [53]. The same approach was used by similar studies in the field of engineering data-intensive applications (e.g., [41, 42]).

The collaboration with Mondragon lasted twelve months. In regular virtual working sessions, the team members discussed design decisions, open issues, and reflected on the experiences they have made. Notes and protocols of these working sessions form the primary source of data for deriving the lessons learned, benefits, and barriers. During project execution, the team followed the action research cycle, which consists of five iterative steps: *Diagnosis*, *Action Planning and Design*, *Action Taking*, *Evaluation*, and *Specifying Learning* [53, 62].

3.1 Diagnosis

The Mondragon Corporation is a large Spanish federation of cooperatives with over 80,000 employees where different research centres, companies, and a university interact to promote new business initiatives. Mondragon is active in several business domains, including industry, finance, retail, and knowledge. Mondragon companies have started a process of digital adaptation to face the great transformation that business, processes, and jobs are experiencing. Collaborations among industrial companies within and outside Mondragon Corporation have increased in recent years. Within these collaborations, pipelines are created in which industrial assets, sensors, and processes produce large amounts of data that need to be collected in different repositories for analysis. The number of companies participating in these initiatives leads to a complex network of distributed data sources and lengthy data management and governance processes.

In this case study, we focus on a production plant at Mondragon where data from several processes is collected. Each process is a sequence of actions involving devices that collaborate in the production of goods. During production, the devices involved in each process generate data in the form of messages related to (1) production parameters measured by the machines involved (e.g., temperature, pressure, etc.), and (2) event data to monitor process performance (e.g., start time, end time, etc.).

Mondragon collects these messages to monitor and evaluate industrial production conditions and process performance. Conditions and performance are satisfactory when parameter values meet the estimated operating requirements or thresholds. However, under those situations, faulty products might appear at any time. Consequently, it is important to determine the cause for defective products and relate it to undesired production conditions or poor process performance. To do so, an anomaly detection step is introduced to the data processing. This step finds irregularities and investigates the root cause of a faulty product. With this data-driven approach to product quality, Mondragon aims to reduce the number of faulty products, the amount of manual quality work, and ultimately achieve the goal of zero-defect manufacturing [47].

While data collection, preparation, and simple data analysis is usually done locally at the plant level, more advanced ML and AI methods often require the assistance of internal or external service providers (see Figure 1). These service providers might be used for several reasons: (1) the technologies and tools for data analysis are not available locally, (2) expert knowledge is required for complex algorithms, or (3) complex data enrichments and integrations are necessary for a more useful result.

To enable the integration of external parties in the data processing, a trusted environment is needed that ensures data is used only by the desired company and under the negotiated conditions for access and use. For Mondragon, a Non-Disclosure Agreement (NDA) was not sufficient to create this level of trust. They aimed to extend the existing legal binding of an NDA with a technical solution that guarantees data are handled in accordance with the agreed policies.

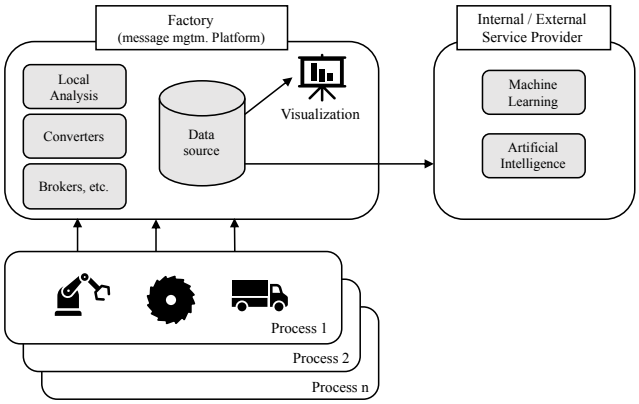

Figure 1: Separation of Data Analysis Tasks at Mondragon

In the case at hand, Mondragon collaborated with a research facility that provides AI services for data quality analysis (see Figure 2). Specifically, it was Mondragon's goal to conduct an outlier and concept drift analysis on data collected from a process involving a press machine. The data quality analysis helps Mondragon validate their data streams and ensure the correctness of consecutive decision making [19, 25]. The research facility uses the data provided by Mondragon to continuously train an AI model that is used for data quality estimation. This model is made available under the same terms and conditions as the data stream. The result is a collaborative, data-sovereign AI pipeline that helps Mondragon leverage their data sets and enables a new business model for the data quality service provider [51].

The action research team that investigated the proposed research questions consisted of three researchers, two practitioners, and four project stakeholders. The researchers and practitioners collaborated on the realization of the desired data-sovereign AI pipeline at Mondragon [53, 62]. Two researchers and the two practitioners formed the core development team. The third researcher provided methodological and architectural guidance. Travel restrictions due to Covid-19 hindered the team from directly working together. To overcome this problem, regular virtual working sessions were used to clarify open questions, for pair programming, and resolving technical and management problems. The project stakeholders included members of Mondragon's senior management and helped the development team prioritize functionalities and mitigate risks.

3.2 Action Planning and Design

Following Petersen et al. [53], the *Action Planning and Design* phase is used to identify and discuss different approaches for problem solving and to choose a suitable approach. In this case, two parties,

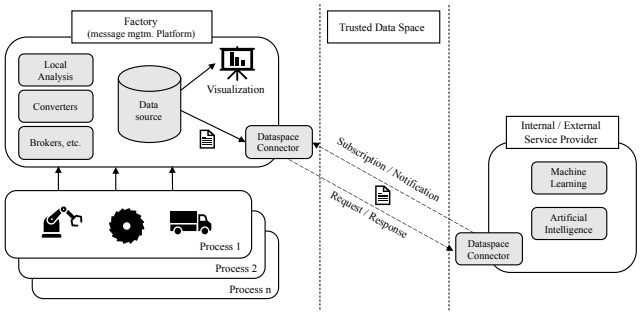

Figure 2: Collaborative and Sovereign Data Analysis at Mondragon

Mondragon and the research facility, collaborated on a federated, collaborative AI pipeline (see Figure 3). Hereby, Mondragon specified the desired model and collected and prepared the data (indicated by the green area). The research facility acted as an AI consultancy and was responsible for training, evaluation, deployment, and monitoring of the AI model used for data quality estimation (blue area). Both parties required data sovereignty for their artifacts, which is the raw data at Mondragon and the trained AI model at the research facility (grey area). Specifically, the partners raised the following four data sovereignty requirements for their respective artifacts: (1) the data can only be used under agreed terms, (2) data access can be revoked at any time, (3) both partners share the same secure execution environment to avoid data loss, and (4) participants can be forced to delete data if necessary.

To realize the desired data-sovereign AI pipeline, the action research team analyzed and compared existing data sovereignty solutions. The IDS principles, as a solution for an implementation of sovereign data processes, appeared to be a good approach to implement the defined requirements. As a concrete implementation of these principles, the action research team selected the DSC [34] as basis for the project. We made this decision for several reasons. The DSC is a reference implementation of an IDS Connector and ensures a sovereign data exchange by following defined standards and rules. As an up-to-date and community-driven open source project, it provides the possibility of continuous exchange, support, and active development. Unlike other open source projects, the DSC is developed and maintained under administration and supported by an official association, the IDSA [7].

With regard to the IDS Reference Architecture Model, the DSC is one of the most advanced connector implementations. It uses state-of-the-art technologies and standards, such as X.509 certificates and a REST API following the RFC 7231 standard [20]. In addition, it implements best practices (e.g., following code style guides, high level of test coverage) and well-known software development patterns (MVC, Factory) to provide high code quality. Thereby, it offers a simple out-of-the-box solution with clearly defined interfaces and a simple deployment. In contrast to other implementations, as presented in Section 2.3, it encapsulates the IDS logic from connected legacy systems and strives for a user-friendly implementation of usage policies and their enforcement.

3.3 Action Taking

The action research cycle continues with the *Action Taking* step in which the implementation of the previously selected approach is described [62]. The implementation requires the realization of three different components. First, both parties need to implement an instance of the *Dataspace Connector (DSC)* [34] to participate in the data space. Second, Mondragon needs to implement a *backend component* that gathers the data and publishes it using the DSC. Third, the research facility needs to offer their AI application for data quality analysis as a service within the IDS data space, a so called *DataApp* [52].

The **Dataspace Connector (DSC)** is an open-source Java application that enables the participation in an IDS-based data space [34]. By using a containerized deployment, it can be easily set up in different system landscapes. The DSC guarantees compliance with agreed data sharing terms in three ways. First, via its REST API, the offered resources are provided with meta information such as data sharing policies (see Figure 4). Before data is shared, a contract negotiation takes place between two connectors to ensure the data sharing follows these policies. Therefore, a potential data consumer either adopts an initial contract offer or provides a counter offer. Its validity is checked by means of syntax, content, and signature. The negotiation concludes with a contract agreement. For increased trust, the IDS Clearing House is included as an attesting third party [52]. Second, to ensure that the containers adhere to the data sharing terms, the IDSA (i.e., the independent trustee) certifies the connector technologies. For this, each DSC must, besides a number of other specifications (see [32]), be equipped with IDS and TLS certificates (both SSL) to be able to share data via a secured communication protocol. Third, the DSC receives consumed data in an internal database and enforces the agreed-on data sharing terms (e.g., deletion after n-times usage) using this database if necessary. Currently, usage control ends at the DSC and is not enforced in other systems. These kinds of usage controls are part of future work (see also Section 4.3). Overall, the DSC attains the necessary level of trust by combining technical and governance measures.

The **Backend Component (Mondragon)** is a data collection, management, and sharing platform based on open-source technologies deployed at Mondragon. The platform consists of three main components: (1) a messaging system based on RabbitMQ, (2) a data flow integration framework implemented with NodeRED, and (3) a data repository constructed with MongoDB. Hereby, the backend collects messages from the processes running in the production plant in JSON format. The messages carry data produced by the devices and assets deployed in the plants. To receive the messages, there are NodeRED data flows subscribed to the RabbitMQ message broker that extracts those messages and locally validates if their values are within the correct thresholds. The messages received are then stored in different MongoDB collections depending on the process they originated from. Based on the nature of a message and the intended data analysis task, the correct AI pipeline is selected. For example, data from a press machine that is intended for external analysis is stored in a corresponding MongoDB collection. Since the interaction with the DSC is REST-based, we collected messages for a time period of 15 minutes and added a unique identifier to the data set. The identifier connects the messages in a data set with the

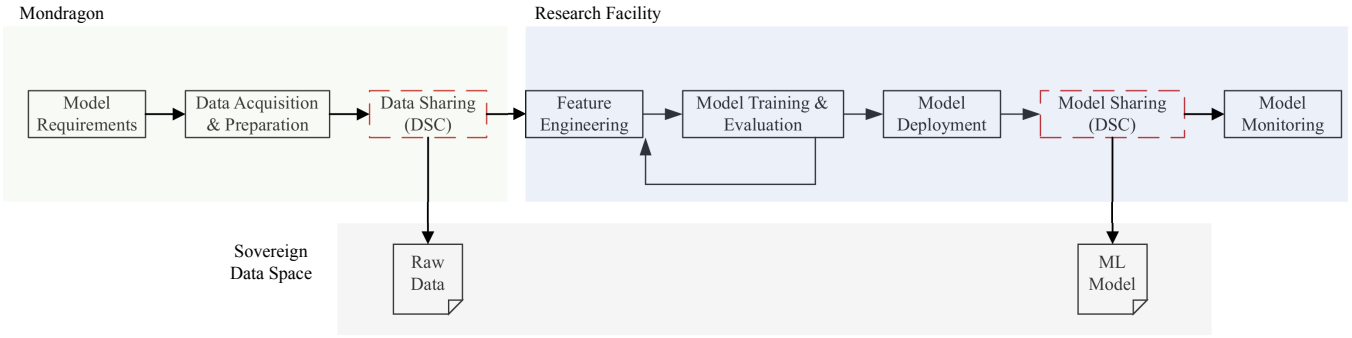

Figure 3: Collaborative Data-Sovereign AI Pipeline, adapted from Amershi et al. [6]

identifier of the AI model created by the external data quality service provider. A typical data set produced this way is a structured JSON file that contains numerous messages with sensor values and is about 1MB in size.

The **DataApp (Research Facility)** contains the AI functionality for data quality analysis as a service. It is realized as a Python application and uses Apache Spark as the analytical framework. For data quality calculation, we combined four different measures that are suitable for sensor-based inputs and cover the quality dimensions accuracy, completeness, and validity [1, 38]. The calculation of data accuracy is two-fold. First, we used an Isolation-Forest algorithm to determine the commonness of outliers in the data set. Second, we assessed a potential concept drift by learning the boundaries and averages of each sensor as an approximation. This approximation can be compared to the approximations of subsequent data sets leading to a measure of concept drift [25]. For completeness, we used a 'No Value' measure that identifies sensors which do not provide data. Finally, we assessed the validity of the data by detecting sensors that stay constant over multiple data sharing cycles.

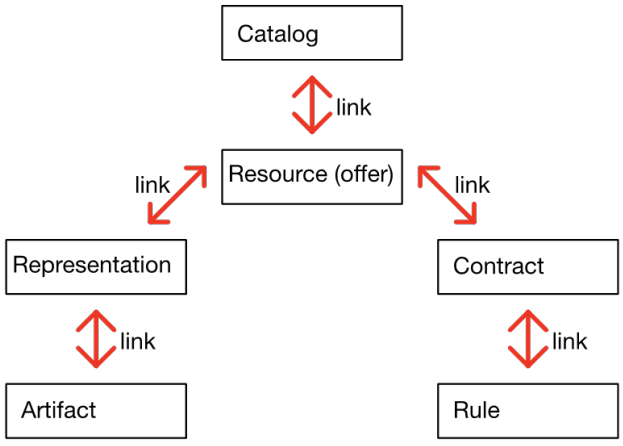

Figure 4: Required Artifacts as Specified by the IDS Information Model [52]

Table 1 summarizes the message flows of the components in the architecture depicted in Figure 2. Hereby, the DSC at Mondragon is referred to as the data source and the one at the research facility as the data sink. A typical cycle time of this process (steps 1 to 7) takes about one minute.

3.4 Evaluation

The *Evaluation* step of the action research cycle measures the effectiveness of the previous *Action Taking* step by conducting focus group discussions, interviews, or questionnaires [53, 62]. To evaluate the implemented data-sovereign AI pipeline at Mondragon, we presented and discussed the final version of our solution with the whole action research team as part of a qualitative focus group discussion. Focus group discussions are well-suited for efficiently gathering relevant data and taking advantage of the group interaction [28]. The meeting included nine participants and lasted for 60 minutes. One of the authors demonstrated the functionality of the AI pipeline, presented the overall architecture, and asked for feedback regarding functional and non-functional characteristics. Furthermore, we asked for potential future developments and missing features.

Overall, we received positive feedback for the AI pipeline from the different stakeholders. Specifically, the participants agreed that our solution is a step in the right direction to enable collaborative AI initiatives with external partners and overcome current data governance and trust issues. This way, our developments help Mondragon to better exploit their available data sets and improve business operations. We included the feedback from this group discussion in our findings, which are presented in Section 4. Regarding the future development of the tool, Mondragon is currently discussing internally how to proceed and what suitable follow-up use cases could be implemented.

3.5 Specifying Learning

In the final step of the action research cycle, the general learnings are identified and formalized based on the previous evaluation [62]. After the evaluation step, the core development team conducted a full-day workshop to reflect on the design decisions, downsides, and success factors they experienced during the twelve-month project. Based on these experiences we formulate generalizable lessons learned, benefits, and barriers. We used different data sources to draw our conclusions, including architecture documents, meeting notes and protocols, and email correspondence [57]. To reach a

Table 1: Message Flows for the Implemented Solution

Step	*Mondragon to Service Provider*	*Service Provider to Mondragon*
1	**Data set Preparation**: Data/factory owner configures system to collect sensor messages for 15 minutes and stores them in a data set. Data owner offers that data set through an API and prepares the IDS Connector offering that resource.	**Data Quality Analysis**: The service provider conducts a data quality analysis from the latest data set using its AI tools. The service provider offers those results through an API and prepares the DSC offering a resource.
2	**Contract Agreement**: Data/factory owner and consumer agree on certificates. Certificates are issued by the IDS certification authority. This is the first step towards a contract agreement for data exchange.	**Contract Agreement**: Data/factory owner and consumer agree on certificates. Certificates are issued by the IDS certification authority. This is the first step towards a contract agreement for data analysis result exchange.
3	**Data Subscription**: The service provider (Data Sink) subscribes to the resource offering the data set through its IDS Connector.	**Data Subscription**: Mondragon subscribes to the resource offering the data quality results through its DSC.
4	**Data Provision**: The application created by the Data owner produces a data set and offers it through its API and the IDS Connector installed in its premises.	**IDS Resource Update**: The Service Provider updates the resource associated to its IDS Connector with the API holding the latest analysis results.
5	**Notification**: The data owner's IDS Connector notifies the data sink's IDS Connector (service provider) about an updated resource.	**Notification**: The service provider notifies the factory owner about new results available (resource update) through the DSC.
6	**Data Request**: The data set is requested by the service provider after a notification is received.	**Result Request**: The analysis results are requested by the factory owner after a notification is received.
7	**Data Sharing**: Usage Policies are checked so that the data set flows only to the IDS Connector complying with the contract. Finally, the data is transferred to the service provider.	**Result Sharing**: Usage Policies are checked so that the results are collected only in the DSC complying with the contracts. Finally, the data quality result is transferred from the service provider to the factory owner.

consensus, we discussed, grouped, and merged different learnings until recurring patterns emerged. We imposed these recurring patterns onto the case at Mondragon and formulated findings that feed into the process of generalization of the class of problem [56]. This means that our findings are not only relevant to Mondragon, but can inform the establishment of data-sovereign AI pipelines in different companies and business domains. We elaborate and discuss these findings in the upcoming Section 4.

4 LESSONS LEARNED, BENEFITS & BARRIERS

After completing the action research project, the project team successfully developed a data-sovereign AI pipeline that supports Mondragon in cooperating with external AI service providers. To structure the presentation of our findings, we used the recently published data challenges model by Gröger [27] for categorization. We noted to what data challenge (i.e., data management, data democratization, and data governance) our findings correspond and how we addressed these challenges.

4.1 Lessons Learned

Based on the aforementioned approach, we formulated ten lessons learned which consist of recurring practices, experiences, and design decisions we made during the project (see Table 2). These lessons learned serve as our response to RQ1.

Need for Data Traceability We soon realized that we needed a solution to link the data sets at Mondragon with their corresponding AI model provided by the research facility. For example, it was difficult for the AI developers at the research facility to resolve mismatches in the data schema due to a lack of data domain knowledge. In accordance with Amershi et al. [6], we found that the 'debuggability' of AI pipelines is an important aspect for its success and usefulness. Especially for complex architectures [6] or safety critical operations [10] such a 'data trail' is vital. In our case, we wanted to be able to track down errors to specific data sets to

Table 2: Lessons Learned for Data-Sovereign AI Pipelines

#	Lessons Learned	Data Challenge [27]
1	Need for Data Traceability	Data Management
2	Need for an Independent Trustee	Data Governance
3	Need for Quality-Driven Data Sharing	Data Democratization
4	Need for a Data Catalog	Data Democratization
5	Need for Real-Time Support	Data Management
6	Need for a Separation of Control and Data Plane	Data Governance
7	Need for Access and Usage Control Enforcement	Data Management
8	Need for Standardization	Data Management
9	Need for a Common Definition of User Roles	Data Governance
10	Need for a Trusted and Secure Deployment Environment	Data Management

identify the source of an error (e.g., data errors, wrong formatting, etc.) and ensure reproducibility. To achieve this goal, both partners extended their data structures and included matching identifiers for corresponding data sets and resulting AI models.

Need for an Independent Trustee One of the biggest obstacles of collaborative AI applications is data privacy [39]. The potential revelation of sensitive or personal information is often seen as too much of a risk, and companies omit the opportunities offered by collaboration. Technical solutions for this problem (e.g., multiparty computation or anonymization) often come with reduced model performance [39]. Another approach is to increase the level of trust between participants and share data on the same terms and conditions, which are guaranteed by an independent trustee. For us, an important motivation to use the DSC was that the IDSA [52] served as such an independent trustee and guaranteed that both parties used the same connector technologies, information models, and terminologies. Technically, this guarantee is realized using IDS certified software components and uniquely identifying the two DSCs using IDS certificates. Without these, data sharing would not be possible within the trusted data space.

Need for Quality-Driven Data Sharing Although Mondragon outsourced the data quality analysis to an external partner in our case, we experienced that an initial data cleaning and preparation step is obligatory before data sharing. Otherwise, simple errors like null values, schema mismatches, or pattern inconsistencies can lead to errors further down the AI pipeline. Data quality is an important aspect of AI pipelines and it is vital to ensure high quality data is shared [6, 27, 42]. Several studies (e.g., [4, 42, 63]) conducted research on realizing data validation in the form of data quality rules, integrity constraints, or 'data tests'. In our case, we specified data schema and pattern validations as 'data tests' to spot changes in the data format and value inconsistencies. This way, we ensured that data and results can be processed by the external party that lacks the necessary data domain knowledge.

Need for a Data Catalog Reusing existing data sets and models has become an important part of AI to reduce duplicate efforts and speed up data science projects [6, 10]. However, reusing data sets requires the ability to find and access suitable data sets. For this, a data catalog, which stores metadata for data artifacts, is a viable solution as it supports the data discoverability and accessibility [27, 44]. In our case, a data catalog was not required because all data sets were known and clearly specified in advance. However, while discussing potential future developments, we found that a data catalog would be beneficial in exploratory AI projects where not all data sets are known a priori. It could, for example, be located at a data space level to bring data providers and consumers together and reduce the usually high efforts associated with data search, access, and preparation [44, 51, 64]. We argue that a data catalog should be considered in more complex AI architectures to facilitate the reuse and sharing of existing data sets.

Need for Real-Time Support During project execution, we were able to confirm the current trend for real-time support on different levels of the AI pipeline. Specifically, the action research team noted that there is a need for supporting data sharing for data streams and a need for the fast deployment and continuous integration of AI models. At Mondragon, we forwarded messages to the AI service provider every 15 minutes, because the DSC REST APIs do not offer real-time support. However, in time-critical use cases this is not an option. Towards this end, Muiruri et al. [44] provide examples of using gRPC (Google Remote Procedure Calls) to realize a real-time, low latency data exchange. With MLOps it has become increasingly popular to connect data science and deployment and overcome typical problems like the imbalance of data scientists and deployment personnel, management of multiple versions of AI models, or training and production skew [37, 43].

Need for a Separation of Control and Data Plane To realize real-time support with a low latency, software has to meet certain technical requirements, as mentioned above. The communication protocols defined by the IDS, for example, can fulfil these requirements only to a certain extent. To enable data sovereignty, this also involves the exchange of additional meta information and the verification of usage policies. For this reason, the data transfer is slower than one without these additional processes. This results in the need to separate the control and data flows in a sovereign system. While steps 1 to 6 from Table 1 are not time-bound and can be implemented using for example an IDS communication protocol, the actual data transfer (step 7) can be implemented using any protocol without losing the added value of data sovereignty. At Mondragon a separation was not required as the usual data transfer was relatively small. It can be important when integrating data-sovereign AI pipelines with very large systems and amounts of data [6].

Need for Access and Usage Control Enforcement What characterizes sovereign systems is not only the possibility to restrict access to the data, but also to define and check its actual usage up to the data consumer. This means that simple authentication and authorization mechanisms through systems such as Keycloak [55] are supplemented by the enforcement of terms of use. The clear definition and control of usage terms helps organizations address prevailing data governance issues [27]. In this context, our data sovereignty solution not only offers the possibility to define usage policies, but also to negotiate them. This increased the level of trust between Mondragon and the research facility and was a major success factor for our project. The negotiated contracts are similar to those specified in data governance protocols. However, the conditions defined with the DSC not only create a general agreement, but can actually be implemented technically. This way, they contribute to the open question of how to implement the usually ill-defined data governance aspects within an organization [27].

Need for Standardization Early in our project, we discovered that there is a lack of experience and standards with regard to data-sovereign AI pipelines (see also Section 2). Although, there are de-facto standards or standardization efforts for data structures (IDS Information Model [9]) and contract negotiation [65], there are still some white spaces. The realization of insular projects with regard to data sovereignty, but also AI in general, hinders the standardization and consequently the proliferation of these concepts [27]. Specifically, we experienced a lack of standardization for development and deployment processes. For example, while it was clear to us what messages the two DSCs needed to exchange (see also Table 1), we had difficulties figuring out how to implement these messages. Other examples include difficulties in integrating our developments with the DSC or problems on connecting our DSC instance with backends. Towards this end, further documentation,

guidelines, and practice reports would be beneficial and help to move from insular projects to a wide application.

Need for a Common Definition of User Roles Data access is a common aspect of the data governance challenge in AI engineering and usually solved by the specification of user roles [27, 64]. In the context of data governance, the literature has defined a number of user roles, such as data steward or data owner, which are widely accepted and used [27, 50]. These roles are usually derived from the structures within a company and then mapped onto their applications and systems. In order for two data sharing systems to negotiate data usage control agreements, the understanding of the included rules must be the same. For example, a data steward might be granted access to an AI model but might have less rights at Mondragon than at the research facility. We experienced this phenomenon when we specified usage controls for the other partner respectively. There needs to be a mutual agreement between the partnering companies on the used user roles. A common definition of these roles, for instance in industry or ecosystem wide data catalogs, would help to achieve this goal.

Need for a Trusted and Secure Environment The recent critical log4j vulnerability [49] showed us once more that software is only as secure as the environment in which it is running. Even if an AI pipeline implements data sovereignty by using encryption and authentication mechanisms and enforcing usage control, this is of limited value if the overall environment has been compromised. Especially, in a collaborative or federated AI initiative, this can be a threat to the mutual trust [10]. Consequently, there is a need to not only implement data sovereignty but also security through all layers, from the software stack to the deployed hardware, e.g., by using Trusted Platform Modules (TPM). Ideally, the data sovereignty component would, hereby, not only guarantee data sovereignty but also enforce system security. The IDS includes "strict container isolation, integrity-protected logging, encryption of all persisted data, protection against accidental misuse by administrators" [32] (p.7) as part of their certification criteria for a trusted IDS Connector. Currently, first developments on technically enforcing these principles are made [34], but there are many opportunities for further developments.

4.2 Benefits

During project execution we observed that a data-sovereign AI pipeline offers several benefits to both partners. We derived these benefits from discussions within the core development team and evaluations with the project stakeholders. Afterwards, we generalized our experiences to formulate common benefits, which serve as a response to RQ2 (see Table 3).

Increased Trust among Participants One of the main benefits of data-sovereign AI pipelines is the increased trust among participants. Having an independent trustee who certifies that all participants are operating under the same terms, conditions, and processes, helps to overcome non-technical obstacles to data sharing that are rooted in a lack of trust. With an increased level of trust, there is a higher willingness for data sharing, which creates a win-win situation for both parties. One team member summarized this benefit as follows:

Table 3: Benefits of Data-Sovereign AI Pipelines

#	Benefits	Data Challenge [27]
1	Increased Trust among Participants	Data Governance
2	Minimization of Data Governance Work	Data Governance
3	Technically Restricted Data Access	Data Management
4	Appliance with Various Technologies and Execution Environments	Data Democratization

Having a guarantee that all participants are certified and using the same technological components and ontologies, makes us feel safe and overall easier to share data sets with partners.

Minimization of Data Governance Work The use of a common framework and information model and the technical enforcement of data usage constraints and sharing contracts reduces the amount of manual data governance work. Considerably, it helps to simplify lengthy processes containing legal clarifications, contract negotiations, and discussions, because all partners are operating on the same basis. As one team member stated:

Sharing data models with external partners is a complicated process including several people and departments, and it can easily take weeks or months until clearance.

Towards this end, a data sovereignty component can help to avoid the establishment of 'data governance anti patterns' [44] and ease data access.

Technically Restricted Data Access In addition to the minimization of data governance work, the technical restriction of data access and the inseparable metadata exchange can help to avoid data management work [27]. The use of a common and machine-readable metadata structure, such as the IDS Information Model [9], enables the technical implementation of this. Guidelines and specifications such as those of the IDS provide a framework and enrich existing pipelines that are based on data sharing. One member of the development team stated the following:

Limiting access to a certain amount simplifies the data management and makes the system more secure by avoiding illicit data access.

Appliance with Various Technologies and Execution Environments Our experiences over the course of this project showed us that data sovereignty is not limited to certain data, technologies, or deployment environments. It can add value in both IoT systems and cloud environments. Moreover, it can integrate new and complex AI processes in a way that does not compromise security and trust in data processing, or result in a modification of existing implementations. It facilitates the reuse and sharing of AI models with other companies and creates new business models [10]. By using complementary software such as the DSC, companies can quickly expand existing systems and integrate them into a data ecosystem without any major effort. Thereby, topics such as data processing and data sovereignty can be easily reconciled. One team member summarized this benefit as follows:

The realization of data sovereignty offers many new opportunities for inter-organizational collaboration to us.

4.3 Barriers

In contrast to the benefits we also observed that there are some barriers to data-sovereign AI pipelines. These are problems, downsides, or limitations of our current solution and provide opportunities for future research and developments. We summarized the derived barriers in Table 4 as a response to RQ3.

Table 4: Barriers to Data-Sovereign AI Pipelines

#	Barriers	Data Challenge [27]
1	Limited Support for Existing Technologies	Data Democratization
2	Potential Performance Issues	Data Management
3	Challenging Implementation of Usage Control	Data Management

Limited Support for Existing Technologies One of the major problems we identified is that current data sovereignty solutions are still in their infancy and it can be difficult to connect them with existing technologies. Especially, when these technologies do not support REST APIs, such as legacy systems, an integration is currently not possible. This imposes limitations on the automated data acquisition, which hinders an adoption of MLOps [37]. To overcome this problem, an intermediary broker (e.g., Apache Kafka) or custom glue code is necessary to connect the incompatible components. Such an intermediary component could be valuable in any case to realize quality-driven data sharing (see also Section 4.1). In our case, Mondragon used NodeRED as an integration framework to connect and distribute messages between different backend systems.

Potential Performance Issues An important conceptual downside of our current solution is that all shared data is transferred through the DSC. This makes sense with regard to the aspect of ensuring data sovereignty, but the additional step can lead to performance issues and an increased latency. At Mondragon, we were able to avoid this issue by transferring only small amounts of data and realizing a non time-critical application. However, for big data and time-critical scenarios the current concept could cause problems. A potential solution for this is a separation of concerns between data transferring and handling usage constraints. However, further technical and conceptual developments are necessary. The EDC project [23] is currently conducting research in this direction.

Challenging Implementation of Usage Control In the previous sections, we elaborated on the need for data sovereignty in the form of common information models, guidelines, and access and usage control. However, for Mondragon and the research facility data sovereignty ends at the DSC. A guaranteed usage control across the whole data lifecycle is very difficult to realize technically [45, 64, 66]. With the IDS Usage Control Language [61] and frameworks such as OPA [17] or LUCON [59], and the implementation of the DSC [34], first steps have been taken. However, how access and usage control along the entire data processing chain can be guaranteed has not yet been comprehensively solved. Even if a system has been implemented securely and trusted at all layers [12] and is integrated into a sovereign ecosystem, the sovereign data will probably leave these systems at some point. It might end up in associated AI applications, existing databases, or other systems that do not have a native usage control implementation. First approaches [14, 23] have addressed this issue, but so far there is no viable solution.

5 CONCLUSION

The guiding objective of our empirical action research case study was to implement an AI pipeline at Mondragon that enables the sovereign data exchange with external AI service providers. To derive generalizable contributions, we reflected and synthesized the experiences and evaluation results of the twelve-month project.

Following Petersen et al. [53], action research is a suitable approach for "transferring research results into practice" (p.61). The generalized findings can be used by other practitioners to implement data sovereignty in AI pipelines in their respective contexts and avoid the emergence of 'data governance anti-patterns' [44]. Furthermore, by presenting our course of actions, we are able to provide insights on the internal perspective of implementing data sovereignty, which can help to raise awareness for sovereign data sharing [15].

From a scientific perspective, our work contributes to the body of knowledge on AI engineering. Specifically, it offers details on a solution to common challenges in federated, distributed AI infrastructures, highlighted by several studies (e.g., [6, 10, 27, 40, 44]). By categorizing our findings in the context of the AI challenges framework by Gröger [27] and highlighting gaps in current solutions, we offer concrete opportunities for future research. Addressing these gaps would help to advance the field and address the prevailing data challenges for AI. Through the continuous interplay of researchers and practitioners, our derived knowledge is both theoretically grounded and practically inspired and constitutes a sound contribution to the scientific community [8, 53].

Our work is subject to several limitations, which are based on the nature of the action research project and the qualitative data analysis [8, 53]. Most importantly, our study faces organizational bias as we conducted our research within a single organization, and is of limited external validity. In other contexts (e.g., smaller companies, different industries, etc.), a solution might look different and lead to other findings. Additionally, the process of reflecting and synthesizing the findings is subjective and other researchers might come to different conclusions.

Further empirical studies in other contexts could help to overcome these limitations and add validity to our findings. Moreover, by investigating further cases, we could advance our findings from general lessons learned to concrete design patterns for data-sovereign AI pipelines. Finally, we plan to address the barriers we identified as part of the further development of our solution to create a more sophisticated data-sovereign AI pipeline.

ACKNOWLEDGMENTS

This research was partly supported by the EU's Horizon 2020 program and the QU4LITY project (GA no. 825030). We also thank Javier Cuenca, Stephan Dübler, Alain Perez, and Ronja Quensel for their support in conducting the research.

REFERENCES

[1] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. 2016. Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment* 9, 12 (2016), 993–1004.

[2] Amir Shayan Ahmadian, Jan Jürjens, and Daniel Strüber. 2018. Extending model-based privacy analysis for the industrial data space by exploiting privacy level agreements. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, New York, NY, USA, 1142–1149. https://doi.org/10.1145/3167132.3167256

[3] Álvaro Alonso, Alejandro Pozo, José Cantera, Francisco de la Vega, and Juan Hierro. 2018. Industrial Data Space Architecture Implementation Using FIWARE. *Sensors* 18, 7 (2018), 22–26. https://doi.org/10.3390/s18072226

[4] Marcel Altendeitering and Tobias Guggenberger. 2021. Designing data quality tools: findings from an action design research project at Boehringer Ingelheim. *29th European Conference on Information Systems* (2021).

[5] Antonello Amadori, Marcel Altendeitering, and Boris Otto. 2020. Challenges of Data Management in Industry 4.0: A Single Case Study of the Material Retrieval Process. In *Business Information Systems*, Witold Abramowicz and Gary Klein (Eds.). Springer International Publishing, Cham, 379–390. https://doi.org/10.1007/978-3-030-53337-3_2

[6] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

[7] International Data Spaces Association. 2022. Members. https://internationaldataspaces.org/we/members/ (Accessed: 08.01.2022).

[8] David E Avison, Francis Lau, Michael D Myers, and Peter Axel Nielsen. 1999. Action research. *Commun. ACM* 42, 1 (1999), 94–97.

[9] Sebastian R. Bader, Jaroslav Pullmann, Christian Mader, Sebastian Tramp, Christoph Quix, Andreas W. Müller, Haydar Akyürek, Matthias Böckmann, Benedikt T. Imbusch, Johannes Lipp, Sandra Geisler, and Christoph Lange. 2020. The International Data Spaces Information Model – An Ontology for Sovereign Exchange of Digital Content. In *International Semantic Web Conference*. Springer, 176–192. https://doi.org/10.1007/978-3-030-62466-8_12

[10] Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic. 2021. Engineering ai systems: A research agenda. In *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*. IGI Global, 1–19. https://doi.org/10.4018/978-1-7998-5101-1.ch001

[11] Arnaud Braud, Gael Fromentoux, Benoit Radier, and Olivier Le Grand. 2021. The Road to European Digital Sovereignty with Gaia-X and IDSA. *IEEE Network* 35, 2 (2021), 4–5. https://doi.org/10.1109/MNET.2021.9387709

[12] Gerd S. Brost, Manuel Huber, Michael Weiß, Mykolai Protsenko, Julian Schütte, and Sascha Wessel. 2018. An Ecosystem and IoT Device Architecture for Building Trust in the Industrial Data Space. In *Proceedings of the 4th ACM Workshop on Cyber-Physical System Security*. ACM, New York, NY, USA, 39–50. https://doi.org/10.1145/3198458.3198459

[13] Fabian Bruckner and Falk Howar. 2021. Utilizing Remote Evaluation for Providing Data Sovereignty in Data-sharing Ecosystems. In *Proceedings of the 54th Hawaii International Conference on System Sciences*. 7005–7014. https://doi.org/10.24251/HICSS.2021.842

[14] Fabian Bruckner, Julia Pampus, and Falk Howar. 2021. A Policy-Agnostic Programming Language for the International Data Spaces. In *Data Management Technologies and Applications*, Slimane Hammoudi, Christoph Quix, and Jorge Bernardino (Eds.). Springer International Publishing, Cham, 172–194. https://doi.org/10.1007/978-3-030-83014-4_9

[15] Arnab Chakrabarti, Christoph Quix, Sandra Geisler, Jaroslav Pullmann, Artur Khromov, and Matthias Jarke. 2018. Goal-Oriented Modelling of Relations and Dependencies in Data Marketplaces.. In *Proceedings of the 11th Inter-national Workshop i* co-located with the 30th International Conference on Advanced Information Systems Engineering*.

[16] European Commission. 2021. A European Strategy for data. https://digital-strategy.ec.europa.eu/en/policies/strategy-data (Accessed: 08.01.2022).

[17] Open Policy Agent contributors. 2022. Policy-based control for cloud native environments. https://www.openpolicyagent.org/ (Accessed: 10.01.2022).

[18] Holger Drees, Dennis O. Kubitza, Johannes Lipp, Sebastian Pretzsch, and Christoph Schlueter Langdon. 2021. Mobility Data Space - First Implementation and Business Opportunities. https://dih.telekom.net/wp-content/uploads/2021/08/ITSWC21_MobiDS_02-00.pdf (Accessed: 08.01.2022).

[19] Lisa Ehrlinger, Verena Haunschmid, Davide Palazzini, and Christian Lettner. 2019. A DaQL to Monitor Data Quality in Machine Learning Applications. In *Database and Expert Systems Applications*, Sven Hartmann, Josef Küng, Sharma Chakravarthy, Gabriele Anderst-Kotsis, A Min Tjoa, and Ismail Khalil (Eds.). Springer International Publishing, Cham, 227–237. https://doi.org/10.1007/978-3-030-27615-7_1

[20] R. Fielding and J. Reschke. 2014. *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. Standard. Internet Engineering Task Force (IETF). https://datatracker.ietf.org/doc/html/rfc7231 (Accessed: 15.01.2022).

[21] e.V. FIWARE Foundation. 2021. FIWARE - Open APIs for Open Minds. https://www.fiware.org/ (Accessed: 10.01.2022).

[22] The Apache Software Foundation. 2022. Apache Camel. https://camel.apache.org/ (Accessed: 19.01.2022).

[23] The Eclipse Foundation. 2022. Eclipse Dataspace Connector. https://github.com/eclipse-dataspaceconnector/DataSpaceConnector (Accessed: 10.01.2022).

[24] Olivier Gallay, Kari Korpela, Niemi Tapio, and Jukka K. Nurminen. 2017. A peer-to-peer platform for decentralized logistics. In *Proceedings of the Hamburg International Conference of Logistics (HICL)*. 19–34. https://doi.org/10.15480/882.1473

[25] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.

[26] Joshua Gelhaar and Boris Otto. 2020. Challenges in the emergence of data ecosystems. In *Proceedings of the 24th Pacific Asia Conference on Information Systems*.

[27] Christoph Gröger. 2021. There is No AI without Data. *Commun. ACM* 64, 11 (2021), 98–108. https://doi.org/10.1145/3448247

[28] Jocelyn A Hollander. 2004. The social contexts of focus groups. *Journal of contemporary ethnography* 33, 5 (2004), 602–637.

[29] Arghavan Hosseinzadeh, Andreas Eitel, and Christian Jung. 2020. A Systematic Approach toward Extracting Technically Enforceable Policies from Data Usage Control Requirements. In *Proceedings of the 6th International Conference on Information Systems Security and Privacy*. 397–405. https://doi.org/10.5220/0008936003970405

[30] Patrik Hummel, Matthias Braun, Max Tretter, and Peter Dabrock. 2021. Data sovereignty: A review. *Big Data & Society* 8, 1 (2021), 1–17. https://doi.org/10.1177/2053951720982012

[31] Fraunhofer IESE. 2022. MY DATA Control Technologies. https://www.mydata-control.de/ (Accessed: 10.01.2022).

[32] International Data Spaces Association. 2019. IDS Certification explained. https://internationaldataspaces.org/wp-content/uploads/IDSA-Position-Paper-IDS-Certification-Explained.pdf (Accessed: 08.01.2022).

[33] Fraunhofer IOSB. 2022. Bauhaus.MobilityLab. https://bauhausmobilitylab.de/en/ (Accessed: 10.01.2022).

[34] Fraunhofer ISST. 2022. Dataspace Connector. https://github.com/International-Data-Spaces-Association/DataspaceConnector (Accessed: 06.01.2022).

[35] Valentina Janev, Maria Esther Vidal, Kemele Endris, and Dea Pujic. 2021. Managing Knowledge in Energy Data Spaces. In *Companion Proceedings of the Web Conference 2021*. ACM, New York, NY, USA, 7–15. https://doi.org/10.1145/3442442.3453541

[36] Matthias Jarke, Boris Otto, and Sudha Ram. 2019. Data Sovereignty and Data Space Ecosystems. *Business and Information Systems Engineering* 61, 5 (2019), 549–550. https://doi.org/10.1007/s12599-019-00614-2

[37] Meenu Mary John, Helena Holmström Olsson, and Jan Bosch. 2021. Towards MLOps: A Framework and Maturity Model. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. 1–8. https://doi.org/10.1109/SEAA53835.2021.00050

[38] Aimad Karkouch, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. 2016. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications* 73 (2016), 57–81.

[39] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60.

[40] Lucy Ellen Lwakatare, Aiswarya Raj, Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic. 2019. A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation. In *Agile Processes in Software Engineering and Extreme Programming*, Philippe Kruchten, Steven Fraser, and François Coallier (Eds.). Springer International Publishing, Cham, 227–243. https://doi.org/10.1007/978-3-030-19034-7_14

[41] Lucy Ellen Lwakatare, Ellinor Rånge, Ivica Crnkovic, and Jan Bosch. 2021. On the Experiences of Adopting Automated Data Validation in an Industrial Machine Learning Project. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 248–257. https://doi.org/10.1109/ICSE-SEIP52600.2021.00034

[42] Aiswarya Raj M, Jan Bosch, Helena Holmström Olsson, and Tian J. Wang. 2020. Towards Automated Detection of Data Pipeline Faults. In *27th Asia-Pacific Software Engineering Conference (APSEC)*. 346–355. https://doi.org/10.1109/APSEC51365.2020.00043

[43] Sasu Mäkinen, Henrik Skogström, Eero Laaksonen, and Tommi Mikkonen. 2021. Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help? *arXiv preprint arXiv:2103.08942* (2021).

[44] Dennis Muiruri, Lucy Ellen Lwakatare, Jukka K. Nurminen, and Tommi Mikkonen. 2021. Practices and Infrastructures for ML Systems An Interview Study in Finnish Organizations. https://doi.org/10.36227/techrxiv.16939192.v2

[45] Andres Munoz-Arcentales, Sonsoles López-Pernas, Alejandro Pozo, Álvaro Alonso, Joaquín Salvachúa, and Gabriel Huecas. 2019. An Architecture for Providing Data Usage and Access Control in Data Sharing Ecosystems. *Procedia Computer Science* 160 (2019), 590–597. https://doi.org/10.1016/j.procs.2019.11.042

[46] Andres Munoz-Arcentales, Sonsoles López-Pernas, Alejandro Pozo, Álvaro Alonso, Joaquín Salvachúa, and Gabriel Huecas. 2020. Data Usage and Access Control in Industrial Data Spaces: Implementation Using FIWARE. *Sustainability* 12, 9 (2020), 38–85. https://doi.org/10.3390/su12093885

[47] Odd Myklebust. 2013. Zero defect manufacturing: a product and plant oriented lifecycle approach. *Procedia CIRP* 12 (2013), 246–251.

[48] Michael Nast, Benjamin Rother, Frank Golatowski, Dirk Timmermann, Jens Leveling, Christian Olms, and Christian Nissen. 2020. Towards an International Data Spaces Connector for the Internet of Things. *2020 16th IEEE International Conference on Factory Communication Systems (WFCS)* (2020), 1–4. https://doi.org/10.1109/WFCS47810.2020.9114503

[49] NIST. 2022. CVE-2021-44228 Detail. https://nvd.nist.gov/vuln/detail/CVE-2021-44228 (Accessed: 10.01.2022).

[50] Boris Otto. 2011. A morphology of the organisation of data governance. In *ECIS 2011 Proceedings*.

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

[52] B. Otto, S. Steinbuss, A. Teuscher, and S. Lohmann. 2019. IDS Reference Architecture Model. https://doi.org/10.5281/zenodo.5105529

[53] Kai Petersen, Cigdem Gencel, Negin Asghari, Dejan Baca, and Stefanie Betz. 2014. Action Research as a Model for Industry-Academia Collaboration in the Software Engineering Context. In *Proceedings of the 2014 International Workshop on Long-Term Industrial Collaboration on Software Engineering* (Vasteras, Sweden) *(WISE '14)*. Association for Computing Machinery, New York, NY, USA, 55–62. https://doi.org/10.1145/2647648.2647656

[54] Sobah Abbas Petersen, Zohreh Pourzolfaghar, Iyas Alloush, Dirk Ahlers, John Krogstie, and Markus Helfert. 2019. Value-Added Services, Virtual Enterprises and Data Spaces Inspired Enterprise Architecture for Smart Cities. In *Collaborative Networks and Digital Transformation*, Luis M. Camarinha-Matos, Hamideh Afsarmanesh, and Dario Antonelli (Eds.). Springer International Publishing, Cham, 393–402.

[55] Inc. Red Hat. 2022. Open Source Identity and Access Management. https://www.keycloak.org/ (Accessed: 16.01.2022).

[56] Lee Peter Ruddin. 2006. You can generalize stupid! Social scientists, Bent Flyvbjerg, and case study methodology. *Qualitative inquiry* 12, 4 (2006), 797–812.

[57] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering* 14, 2 (2009), 131–164. https://doi.org/10.1007/s10664-008-9102-8

[58] David Sarabia-Jacome, Ignacio Lacalle, Carlos E. Palau, and Manuel Esteve. 2019. Enabling Industrial Data Space Architecture for Seaport Scenario. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. IEEE, IEEE, 101–106. https://doi.org/10.1109/WF-IoT.2019.8767216

[59] Julian Schütte and Gerd S. Brost. 2018. LUCON: Data Flow Control for Message-Based IoT Systems. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, IEEE, 289–299. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00052

[60] Julian Schütte, Gerd S. Brost, and Sascha Wessel. 2018. Der Trusted Connector im Industrial Data Space. *Fraunhofer-Publication of Fraunhofer Institute for Applied and Integrated Security, Garching* (2018). arXiv:1804.09442 http://arxiv.org/abs/1804.09442

[61] S. Steinbuss et al. 2021. Usage Control in the International Data Spaces. https://doi.org/10.5281/zenodo.5675884 (Accessed: 15.01.2022).

[62] Gerald I Susman and Roger D Evered. 1978. An assessment of the scientific merits of action research. *Administrative science quarterly* (1978), 582–603.

[63] Arun Swami, Sriram Vasudevan, and Joojay Huyn. 2020. Data Sentinel: A Declarative Production-Scale Data Validation Platform. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. 1579–1590. https://doi.org/10.1109/ICDE48307.2020.00140

[64] Daniel Tebernum., Marcel Altendeitering., and Falk Howar. 2021. DERM: A Reference Model for Data Engineering. In *Proceedings of the 10th International Conference on Data Science, Technology and Applications - DATA,*. INSTICC, SciTePress, 165–175. https://doi.org/10.5220/0010517301650175

[65] World Wide Web Consortium (W3C). 2018. ODRL Information Model 2.2. https://www.w3.org/TR/odrl-model/ (Accessed: 16.01.2022).

[66] Johannes Zrenner, Frederik O. Möller, Christian Jung, Andreas Eitel, and Boris Otto. 2019. Usage control architecture options for data sovereignty in business ecosystems. *Journal of Enterprise Information Management* 32, 3 (2019), 477–495. https://doi.org/10.1108/JEIM-03-2018-0058