Agriculture Data Space for Sovereign Data Sharing and Semantic Integration

Roberto García[*][†]and Rosa Gil[‡]

June 7, 2024

Abstract

Following the original spirit of the ”Data Spaces” concept about lowering entry barriers to data sharing by postponing data integration to when it is needed, a data space Proof of Concept has been implemented using the Pontus-X components which also guarantees data sovereignty by design. Different use cases from the agriculture domain have been deployed on this data space and the experience shows that it is possible to implement semantic integration of data generated at an experimental pig farm following a ”pay-as-you-go” approach, while guaranteeing that consumers of that data benefit from it but do not get access to the original data, avoiding thus the risks of it being copied and escaping their control.

Agriculture; Data Space; Data Sovereignty; Data Integration; Semantic Mapping

1 Introduction

The ”Data Spaces” concept was first proposed in 2005 as a shift from central databases to storing data at the source [2]. Data spaces were introduced mainly as a data integration approach, not requiring a physical integration of the data, leaving the data stored at the source, decentralized, and not requiring a common database schema. Integration is achieved, later, when needed, on a semantic level using shared vocabularies. This approach lowers data-sharing entry barriers as providers are not required to do the integration before sharing the data.

Besides this original technological definition of data spaces, their focus on decentralisation has been embraced by current initiatives like the European Union Strategy for Data[1], which set out the path to the creation of Common European Data Spaces[2] in several strategic fields, like health, agriculture, manufacturing, energy, mobility, financial, public administration, skills, media or cultural heritage.

The EU’s interest in data spaces’ decentralisation is motivated by a strategy geared towards achieving data sovereignty, based on open infrastructures for data exchange whose participants are aware and in control of the data they produce and consume, and the services involved. Therefore, institutions, organizations and even individuals are provided data sovereignty through enforceable policies.

Overall, beyond the EU’s data strategy and considering data sharing worldwide, data sovereignty might be a tool to mitigate reluctance to share data while avoiding dominant players and data silos [3], especially if also supported by participation incentives and shared governance where all stakeholders get a say in how data is controlled.

This was our starting point for developing a Proof of Concept (PoC) of a data space in agriculture. First of all, following the original ”Data Spaces” spirit, do not require participants to make the data integration effort beforehand. Thus, data providers can directly make their data available

[*] Computer Engineering and Digital Design Department, Universitat de Lleida, Lleida, Spain. E-mail: roberto.garcia@udl.cat

[†] Corresponding author

[‡] Computer Engineering and Digital Design Department, Universitat de Lleida, Lleida, Spain. E-mail: rosamaria.gil@udl.cat

[1] https://digital-strategy.ec.europa.eu/en/policies/strategy-data

[2] https://digital-strategy.ec.europa.eu/en/policies/data-spaces

through the data space, which features semantic interoperability mechanisms "as needed". Second, to do so using a technological framework guaranteeing data sovereignty by design, as a key mechanism to reduce concerns about data sharing.

Section 2 presents the proposed approaches and tools we used to develop this PoC agriculture data space. Then, Section 3 shows our preliminary results for a particular data-sharing scenario in the pig sector, where semantic integration has been implemented while guaranteeing data sovereignty. Finally, Section 4 presents the conclusions and the future work.

2 Materials and Methods

To develop a Proof of Concept of an agriculture data space providing on-demand semantic integration while guaranteeing sovereignty, we followed the "Pay-as-you-go"[3] approach proposed by Curry et al. [1]. Following the original "Data Spaces" spirit [2], it provides flexibility by reducing the initial cost and barriers to joining the data space. At the minimum level, a data source just needs to be made available as it is within a data space.

Over time, the level of integration can be improved incrementally on a per-need basis using semantic mapping tools like the R2RML[4] W3C standard for expressing customized mappings from relational databases to RDF[5] datasets. By using semantic graph data models combined with formalised vocabularies and ontologies, data semantics are made explicit and it is possible to make it easier to find, access, integrate and reuse in line with the FAIR principles [5].

The vocabularies and ontologies used in a data space, following reference data space architecture models like IDS RAM 4.0[6], are managed by one of its core data space components called the “Vocabulary Hub”. In our case, we are using the AgroPortal ontologies repository [4] to facilitate the reuse of the main ontologies in the agriculture domain.

For sovereignty, we explored existing data space implementations. Most of them provide sovereignty using policy languages that capture the usage conditions defined by the data provider. However, current implementations do not guarantee usage enforcement yet. Though data space connectors implementing data exchange do so after check-in compliance with the usage policy, there are potential risks derived from the storage infrastructure used to store the data, the applications finally using that data or the systems where those applications run. There is no guarantee that the data moved from provider to consumer cannot be copied and re-shared beyond the initial conditions.

To our knowledge, the only readily available solution to implement a data space PoC capable of guaranteeing data sovereignty by design is the one used by the Pontus-X ecosystem[7], which is based on components by OceanEnterprise[8] and deltaDAO[9]. This solution incorporates smart contracts and decentralized technologies to govern data access, storage, and sharing.

Smart contracts are automatically executed if the consumer and service provider sign an agreement and when the conditions are met. Usage restrictions can be part of these contracts. Moreover, audit trails are enabled by logging all transactions on the blockchain. This alone might facilitate checking a posterior if the consumer has complied with the agreed conditions, but it is not enough to provide sovereignty by design.

To do so, the technology stack used by the Pontus-X ecosystem supports data access as well as running computations on data without directly accessing them. Following this compute-to-data approach, data processing services are executed where the data is stored or in a trusted environment provided by an intermediary. Thus, data providers keep control of the data as the consumers do not get access to a copy of the data, just to the results. The next section presents the results of our agriculture data space PoC using the previous methods and tools.

3 Results

The proposed approach has been evaluated through an agriculture data space PoC based on the Pontus-X ecosystem and components. Different data-sharing use cases have been deployed in the data space related to data and services for the digitalisation and the application of Artificial

[3] https://dataspaces.info/pay-as-you-go-services
[4] https://www.w3.org/TR/r2rml/
[5] https://www.w3.org/TR/rdf-concepts/
[6] https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4
[7] https://docs.pontus-x.eu/
[8] https://www.oceanenterprise.io
[9] https://www.delta-dao.com

Intelligence in the pig sector, as illustrated in Figure 1. The data space is available online[10] and has been used by the Centre of Swine Studies of Catalonia (CEP) experimental pig farm to share data collected by precision feeding machines, farm environmental conditions sensors and pig pen cameras.

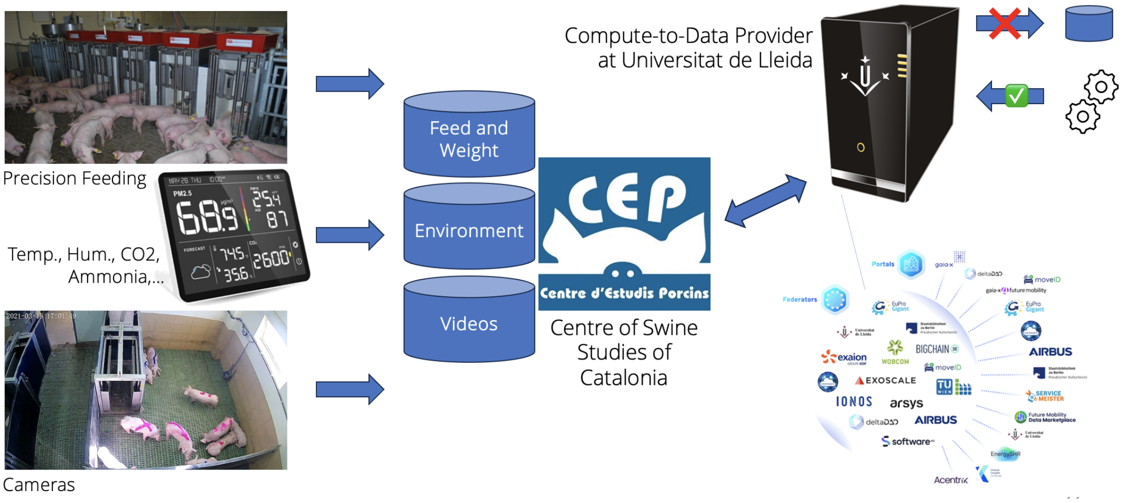

Figure 1: Enter Caption

The data space has used the underlying Pontus-X ecosystem components (blockchain, smart contracts, metadata caches...) and a dedicated Ocean Protocol Provider enabling the compute-to-data functionality. This way, it was possible to guarantee to the experimental pig farm that their data would remain under their control. It is just copied to the trusted environment of the compute-to-data provider, where the data processing services are also deployed. These services are also made available through the data space and packaged as containers. After their execution on the data, both are destroyed and only the outputs are available for download by the consumer.

The use case that best illustrates how the data space provides data sovereignty by design is the one about the computation of animal well-being metrics. The data provider registers into the data space sequences of images from video surveillance of one of its pens. The images can be used by an animal well-being assessment service also available through the data space. It performs automatic image segmentation and tracking of pig movements. Additionally, it also monitors the visits of pigs to defined areas of interest like the automatic feeding machine or the waterer bowl. This allows for the automatic generation of metrics that can be used for animal well-being assessment.

Data sovereignty is guaranteed by design. The algorithm visits the image sequence inside the compute-to-data provider, where they are analysed, and only the computed metrics can be downloaded by the consumer. Consequently, there is no leakage of any image from inside the farm.

Another use case that has been implemented illustrates semantic integration following the ”pay-as-you-go” approach while preserving data sovereignty. Instead of requiring that publishers integrate the data based on existing schemas, which causes a significant upfront overhead, this ”on-demand” approach favours an incremental approach. This way, entry barriers are lowered and it becomes easier to share data.

Moreover, data sovereignty is guaranteed by design by preventing consumers from getting copies of the data and all its processing at the compute-to-data provider. The data to be mapped to semantic form does not leave the compute-to-data environment. It is processed and then stored using a graph store, currently Fuseki, that stays inside that environment and can just be reached from within. This way, it remains under the control of the data provider, the Centre of Swine Studies of Catalonia (CEP).

Later, CEP can decide to grant access to trusted data processing services to visit the compute-to-data environment and slice the Knowledge Graph to extract the semantically integrated data relevant to their computations. Also, in this case, data sovereignty is guaranteed as just computation results, like aggregations or AI-trained models, can leave the compute-to-data environment, not the original data or subsets of it.

As mentioned, CEP can directly share its data using their existing data schemas. For instance, the data produced by one of its precision-feeding machines registers the amount of feed consumed

[10] https://dataspace.angliru.udl.cat

and weight each time the pig uses the machine. This kind of dataset has the following structure based on 7 columns:

- Pen Number: the pen the pig is located in.
- Animal ID: the pig identifier.
- Date: the date the feeding data is about.
- Time: the time pig feeding happened for the given date.
- Duration (s): the duration of the feeding event.
- Feed (g): the amount of food provided, measured in grams.
- Animal Weight (g): the pig weight, measured by the feeding machine during each feeding event.

The data can be shared as is through the data space and, later, any interested party might develop a data mapping service that converts it into an easier-to-integrate format, like RDF based on existing ontologies. There might be different incentives for the development of this mapping afterwards, including the monetization mechanisms provided by the Pontus-X for datasets and data processing services like semantic data mappers.

We have implemented a generic mapper that is already available through the data space based on RMLMapper[11]. It can be configured with different mappings, defined using YARRRML[12] to map specific input tabular data schemas to W3C RDF semantic data based on well-established vocabularies and ontologies that facilitate data integration, even across use cases and activity sectors.

For instance, the mapping defined for files based on the precision-feeding schema previously presented generates RDF data based on the Smart Applications REFerence (SAREF) ontology[13], a vocabulary supported by the European Telecommunications Standards Institute (ETSI) that facilitates data integration in the smart applications domain.

For instance, for the data in Table 1, the mapping generates the RDF represented in graph form in Figure 2, which captures the semantics of the data making them explicit and easier to integrate with other data sources. For instance, units of measure, what is being measured, which of its properties, etc.

Pen Number	Animal ID	Date	Time	Duration (s)	Feed (g)	Animal Weight (g)
4	982091062894196	2021-03-17	10:46	50	14	16500

Table 1: Original tabular data generated by one of the precision-feeding machines

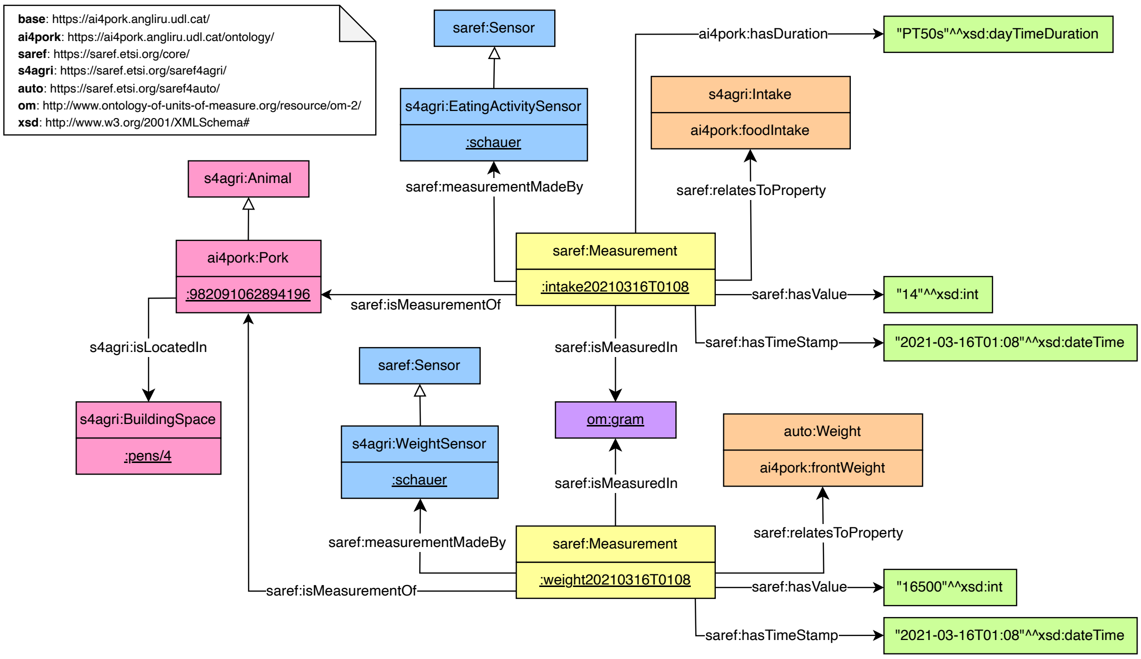

Figure 2: Feed tabular data mapped to RDF and using vocabularies like the SAREF ontology

[11] https://github.com/RMLio/rmlmapper-java

[12] https://rml.io/yarrrml/

[13] https://saref.etsi.org/

4 Conclusions

From our experience with the agriculture data space PoC, we have been able to validate that it is possible to implement data-sharing initiatives capable of guaranteeing data sovereignty by design. We have used currently available data space building blocks implementations available as part of the Pontus-X ecosystem. They make it possible to guarantee data providers that they remain in control of the data they make available through the data space, so the reuse conditions can be enforced.

This is implemented using a compute-to-data approach, where the consumers do not get access to a copy of the provided data. Following this approach, the data remains on provider premises and is just copied to a trusted compute-to-data provider where data processing is performed. The consumer gets access just to the result of that computation, which is checked to guarantee that it does not contain complete or partial copies of the data. The results might contain aggregations of the original data or AI models trained with it.

Our results also show that it is possible to implement semantic integration of the data made available through the data space following a "pay-as-you-go" approach while guaranteeing data sovereignty. The data space PoC has been used by an experimental pig farm to directly share data from their precision feeding machines, implementing data mappers as processing services also made available through the data space. Semantic integration is thus performed later as needed and can be incentivised using the monetization mechanisms available in Pontus-X.

Moreover, semantic integration is performed in a way so that consumers can benefit from the added value of the semantically integrated data, which might combine multiple data sources, without getting access to the original or integrated data. Thus, data sovereignty is guaranteed to data providers, which do not risk their data being copied and reshared beyond their control and the agreed usage conditions.

5 Acknowledgments

This work has been supported by projects “ANGRU: Applying kNowledge Graphs to research data ReUsability” (PID2020-117912RB-C22) and “AI4Pork: Artificial Intelligence in the digital transition of the pig sector” (TED2021-130829B-100), both funded by MCIN/AEI/10.13039/501100011033 and, the latter, also by the European Union NextGenerationEU/PRTR. This research also benefits from funding from the European Union under grant Digital Europe 101100622, “AgrifoodTEF: The European Testing and Experimentation Facilities for Agrifood Innovation”.

References

[1] Curry, E., Derguech, W., Hasan, S., Kouroupetroglou, C., and ul Hassan, U. A real-time linked dataspace for the internet of things: Enabling “pay-as-you-go” data management in smart environments. *Future Generation Computer Systems 90* (Jan 2019), 405–422.

[2] Franklin, M., Halevy, A., and Maier, D. From databases to dataspaces: a new abstraction for information management. *ACM SIGMOD Record 34*, 4 (Dec 2005), 27–33.

[3] Hummel, P., Braun, M., Augsberg, S., and Dabrock, P. Sovereignty and data sharing. *ITU Journal: ICT Discoveries 2* (2018).

[4] Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M. A., Pesce, V., and Larmande, P. Agroportal: A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture 144* (2018), 126–143.

[5] Top, J., Janssen, S., Boogaard, H., Knapen, R., and Şimşek Şenel, G. Cultivating FAIR principles for agri-food data. *Computers and Electronics in Agriculture 196* (May 2022), 106909.