Data Sovereignty and the Internet of Production

Matthias Jarke (✉)
Informatik 5, RWTH Aachen and Fraunhofer FIT,
Ahornstr. 55, 52074 Aachen, Germany
jarke@dbis.rwth-aachen.de

Abstract. While the privacy of personal data has captured great attention in the public debate, resulting, e.g., in the European GDPR guideline, the sovereignty of knowledge-intensive small and medium enterprises concerning the usage of their own data in the presence of dominant data-hungry players in the Internet needs more investigation. In Europe, even the legal concept of data ownership is unclear. We reflect on requirements analyses, reference architectures and solution concepts pursued by the International Data Spaces Initiative to address these issues. The second part will more deeply explore our current interdisciplinary research in a visionary “Internet of Production” with 27 research groups from production and materials engineering, computer science, business and social sciences. In this setting, massive amounts of heterogeneous data must be exchanged and analyzed across organizational and disciplinary boundaries, throughout the lifecycle from (re-)engineering, to production, usage and recycling, under hard resource and time constraints. A shared metaphor, borrowed from Plato’s famous Cave Allegory, serves as the core modeling and data management approach from conceptual, logical, physical, and business perspectives.

Keywords: Data sovereignty · Data spaces · Digital shadows · Internet of Production

1 Introduction

The term “data sovereignty” is hotly debated in political, industrial, and privacy communities. Politicians understand sovereignty as national sovereignty over data in their territory, when it comes to the jurisdiction over the use of big data by the big international players.

One might think that data industries dislike the idea because – in whatever definition – it limits their opportunities to exploit “data as the new oil”. However, some of them employ the vision of data sovereignty of citizens as a weapon to abolish mandatory data privacy rules as limiting customer sovereignty by viewing them as people in need of protection in an uneven struggle for data ownership. For exactly this reason, privacy proponents criticize data sovereignty as a tricky buzzword by the data industry, aiming to undermine the principles of real self-determination and data thriftiness (capturing only the minimal data necessary for a specified need) found in
many privacy laws. The European GDPR regulation follows this argumentation to some degree by clearly specifying that you are the owner of all personal data about yourself.

Surprising to most participants, the well-known Göttingen-based law professor Gerald Spindler, one of the GDPR authors, pointed out at a recent Dagstuhl Seminar on Data Sovereignty (Cappiello et al. 2019) that this personal data ownership is the only formal concept of data ownership that legally exists in Europe. In particular, the huge group of knowledge-intensive small and medium enterprises (SMEs) or even larger user industries in Europe are lacking coherent legal, technical, and organizational concepts how to protect their data- and model-based knowledge in the globalized industrial ecosystems.

In late 2014, we introduced the idea to extend the concept of personal data spaces (Halevy et al. 2006) to the inter-organizational setting by introducing the idea of *Industrial Data Spaces* as the kernel of platforms in which specific industrial ecosystems could organize their cooperation in a data-sovereign manner (Jarke 2017; Jarke and Quix 2017). The idea was quickly taken up by European industry and political leaders. Since 2015, a number of large-scale German and EU projects have defined requirements (Otto and Jarke 2019). Via numerous use case experiments, the International Data Space (IDS) Association with currently roughly 100 corporate members worldwide has evolved, and agreed on a reference architecture now already in version 3 (Otto et al. 2019). Section 2 gives a brief overview of this reference architecture, its philosophy of “alliance-driven data ecosystems”, and a few of the technical contributions required to make it operational.

As recently pointed out by Loucopoulos et al. (2019), the production sector offers particularly complex challenges to such a setting due to the heterogeneity of its data and mathematical models, the structural and material complexity of many products, the globalized supply chains, and the international competition. Funded by the German “Excellence Competition 2019”, an interdisciplinary group of researchers at RWTH Aachen University therefore started a 7-year Excellence Cluster “Internet of Production” aiming to address these challenges in a coherent manner. Section 3 presents an overview of key concepts and points to ongoing work on specific research challenges.

## 2 Alliance-Driven Ecosystems and the International Data Space

### 2.1 Keystone-Driven vs. Alliance-Driven Data Ecosystems

Several of the most valuable firms worldwide create value no longer by producing their own output but by orchestrating the output of others. Following modern versions of early medieval port cities and more recently phone companies, they do this by creating network effects by creating platforms which serve as two-sided or multi-sided markets (Gawer 2014).

In the best-known cases within the software industries, such as Apple, Amazon, or Facebook, but also domain-specific ones like Uber or Flixbus, there is a keystone player defining and running the platform. The typical strategy here is a very high early marketing investment to gain as many early adopters as possible, thus achieving
quickly a dominant market position and being able to exploit extremely rich data sets as a basis for analytics, advertising, or economies of logistics. Design requirements engineering for this kind of platforms was already discussed in (Jarke et al. 2011).

More recently, however, driven by the goals of exploiting the opportunities of platforms but also preserving organizational data sovereignty, we are seeing the appearance of platform-based ecosystems organized and governed by alliances of cooperating players. Examples in Europe include a smart farming data ecosystem initiated by the German-based farm equipment producer Claas together with farming and seed-producing partners, which was recently joined by Claas’ fiercest competitor John Deere.

Another example is an ongoing effort by VDV, the German organization of regional public transport organization, to set up an alliance-driven data ecosystem for intermodal traffic advice, ticketing, and exception handling in competition to efforts by big keystone players such as Flixbus, Deutsche Bahn, or even Google Maps based on Aachen’s Mobility Broker metadata harmonization approach (Beutel et al. 2014).

Beyond the core importance of such a domain-specific meta model, the creation of platform business models is a key success factor. Yoo et al. (2010) already pointed out that, in contrast to traditional business model approaches, the “currency” of multi-sided markets can be an exchange of services rather than a purely financial one. In Pfeiffer et al. (2017), we therefore developed a business model development approach based on their service-dominant business logic and validated it in this intermodal travel ecosystem (cf. Fig. 1). In Otto and Jarke (2019), we employed a literature analysis and elaborate focus groups with industrial partners from the IDS Association in order to identify the main commonalities and differences (Table 1), extending an earlier discussion of platform evolution by Tiwana et al. (2010).
2.2 The IDS Reference Architecture Model

In this subsection, we summarize some of the fundamental design decisions of the International Data Space approach to alliance-driven data ecosystems design. The full description with, e.g., a complete information model of the approach can be found in (Otto et al. 2019) from which the figures in this section have been excerpted.

With its focus on sovereign and secure data exchange, the IDS Architecture takes up aspects of several other well-known architectural patterns for data exchange, as shown in Fig. 2. It is different from the data lake architectures employed by most keystone-driven data platforms which emphasize rapid loading and pay-as-you-go data integration and knowledge extractions, but does embed such functionalities as service offerings whose usage, however, can be limited with enforced and monitored usage policies. On the other side, blockchains can be considered one extreme of such enforcements in a decentralized setting aiming at full transparency and provenance tracking, whereas the IDS architecture emphasizes the sovereign definition of usage policies by the individual players, plus agreed policies for a Data Space.

### Table 1. Alliance and keystone-driven ecosystem design (Otto and Jarke 2019)

| Theoretical concept        | Keystone-driven                                      | Alliance-driven                                                |
|----------------------------|------------------------------------------------------|---------------------------------------------------------------|
| Platform architecture      | Architecture determined by goals of keystone firm    | Architecture determined by shared interest of multiple owners (leading to decentralized data storage, for example) |
| Platform boundary resource | Mainly technical boundary resources (APIs, SDKs etc.), supported by “social” boundary resources (e.g. training for developers) | Data as a boundary resource of “dual” nature, i.e. requiring both technical processing and functional use; many social boundary resources, such as working groups, task forces |
| Platform design            | Core developed by platform owner, then extended by complementors | Consensus oriented design process with focus on “common denominator” |
| Platform ecosystem stages  | 1) Innovation, 2) Adoption, 3) Scaling               | 1) Adoption, 2) Innovation, 3) Scaling                        |
| Ecosystem governance       | Start with limited number of sides and limited options for interaction between them, then increase number of sides and options for interaction | Start with complex ecosystem (i.e. multi-stakeholder setting), then reduce to core ecosystem and extend it later on depending on roll-out requirements |
| Regulatory instruments     | Mainly pricing instruments, accompanied by non-pricing instruments | Dominated by non-pricing instruments; integration of pricing instruments scheduled for scaling phase; data governance |
Membership of an actor (organizational data management entity, cloud, or service provider) is established by two core technical elements, as illustrated in Fig. 3: firstly, the data exchange (only) under agreed Usage Policies (shown as print-like IDS boxes on the links in the figure), and secondly by more or less “trusted” IDS Connectors. These combine aspects of traditional wrapper-mediators for data integration with trust guarantees provided by a hierarchy of simple to very elaborate security mechanisms.

Within the IDS projects, at least four different usage policy enforcement strategies have been developed (Eitel et al. 2017), all accessible via the conceptual model of the ODRL (Open Digital Rights Language) accepted by W3C in 2015. The usage control specifications support model-based code generation and execution based on earlier work by Pretschner et al. (2005). Figure 4 shows how security elements for policy definition (PDP) and policy management (PMP) in the linkage between connectors interact with policy execution points (PEP) in the IDS Connectors from which they can be propagated even to certain specific data items within the protected internal sphere of a company.
In Fig. 1, we referred to the service-dominant business logic underlying most alliance-driven data ecosystems including the IDS. Obviously, as in other trust-intensive inter-organizational settings (Gans et al. 2003), the individual actors and linkages should be carefully defined at the strategic level, for which conceptual modeling techniques such as i* strategic dependencies (Yu 2011) are the obvious candidates. The analysis summarized in (Otto and Jarke 2019) has therefore led to the inclusion of an i*-inspired actor dependency network of important roles and their (task) dependencies (Fig. 5). In the Reference Architecture Model version 3.0 report (Otto et al. 2019), this is further elaborated with business process model patterns for the individual tasks, and governance mechanisms for the organizational infrastructure underlying data ecosystem set-up and operation.

Fig. 4. Usage policy definition, management and execution in the link between IDS subsystems

Fig. 5. Actor-dependency network in the IDS Reference Architecture
3 Towards an “Internet of Production”

Typical IDS use cases so far have been relatively limited in their organizational and technical complexity. A number of more ambitious and socially complex variants, such as Medical Data Spaces for cross-clinic medical and biomedical research have started and are accelerated by the demand caused by the CoVid-19 crisis.

However, probably the most complex application domain tackled so far is production engineering, the subject of our DFG-funded Excellence Cluster “Internet of Production”. In this 7-year effort, 27 research groups from production and materials engineering, computer science, business and social sciences cooperate to study not just the sovereign data exchange addressed by the IDS Architecture in a fully globalized setting, but also the question of how to communicate between model- and data-driven approaches of vastly different disciplines and scales. In this setting, massive amounts of heterogeneous data must be exchanged and analyzed across organizational and disciplinary boundaries, throughout the lifecycle from (re-)engineering, to production, usage and recycling, under hard resource and time constraints.

Figure 6 illustrates this complexity with a few of the different kinds of data, but most importantly with three different lifecycles that have traditionally hardly communicated in the nowadays necessary speed of feedback cycles among them.

As one use case, we are considering the introduction of low-cost electric vehicles to the market. The engineering of such completely new cars requires numerous different specialties and supplier companies to collaborate and exchange data worldwide. To be financially viable, their production will take place in many small factories which can already be profitable with, say, 2,000 cars a year, rather than the traditional 200,000.
But this raises the question how the many small factories all over the world can exchange best practices, and provide feedback to the development engineers for perhaps culture-specific design improvements. Last not least, only the usage experience with buying and operating the novel vehicles will really show what works, what is attractive for customers, what are their ideas for improving the design but perhaps also the production process of the cars. And all of this must happen in an iterative improvement cycle which runs at least 5–10 times the speed of traditional automotive innovation, in order to have a successful vehicle before the venture capital runs out.

In addition to the challenges mentioned in Sect. 2, the computer science perspective on Fig. 6 yields extreme syntactic and semantic heterogeneity combined with very high data volume, but often also highly challenging real-time velocity requirements, and a wide variety of model-driven and data-driven approaches. Thus, all the V’s of Big Data are present in this setting in addition to sovereign data exchange.

Building a complete Digital Twin for such a setting appears hopeless. We need a new kind of data abstraction with the following properties:

- Support for model-driven as well as data-driven approaches, and combinations thereof, across a wide range of different scientific disciplines.
- Relatively small size, in order to be easily movable anywhere between cloud and edge computing.
- Executable according to almost any real-time demand, with reasonable results on at least the most important perspective to the problem at hand.
- Suitable as valorized objects in an IDS-like controlled data exchange ecosystem.

In other words, we must find a common understanding addressing the open question what are actually the objects moving around in an Industrial Data Space.

![Plato’s Cave Allegory](image)

**Fig. 7.** Plato’s Cave Allegory

The intuition for our solution to this question came from an unexpected corner, Greek philosophy. In its famous Allegory of the Cave illustrated in Fig. 7, Plato
(ca. 400 B.C.) showed the limits of human knowledge by sketching a scenery in which humans are fixed in a cave such that they can only see the shadows of things happening in the outside world cast by natural light or by fires lit behind the phenomena (it is funny to note that he even invented the concept of Fake News using human-made artefacts instead of real-world objects between the fire and the shadow).

Anyway, the shadows are obviously highly simplified data-driven real-time enabled abstractions which are, however, created under specific illuminations (=models) such as sunlight or fire. We therefore named our core abstraction *Digital Shadows*, a suitably compact result of combining context-specific simplified models and data analytics. Formally, Digital Shadows can be seen as a generalization of the view concept in databases where the base data as well as the (partially materialized) views are highly dynamic and complex objects.

Once we had invented this abstraction, we found that there already exist quite a number of useful Digital Shadow examples, among them quite well-known ones like the combinations of shortest path algorithms and real-time phone location data shown in navigation systems like TomTom, the flexible combination of Petri net process modeling with event mining algorithms in process mining (Van der Aalst 2011), or abstractions used in real-time vision systems for autonomous driving or sports training.

Within the Excellence Cluster “Internet of Things”, we have been demonstrating the usefulness on complex examples from machine tooling control with extreme real-time constraints, optimizing the energy-intensive step of hot rolling ubiquitous in steel-based production chains, plastics engineering, and laser-based additive manufacturing.

Rather than going into details here, we point to some companion papers elaborating specific issues including the Digital Shadow concept itself and some early validation experiments (Liebenberg and Jarke 2020), the design of an initial physical infrastructure emphasizing the dynamic positioning and secure movement of Digital Shadows as well as base data within a novel infrastructure concept (Pennekamp et al. 2019), logical foundations of correct and rapid data integration in connectors or data lakes (Hai et al. 2019), model-driven development (Dalibor et al. 2020) and the creation of a RDF-based metadata infrastructure called FactDag (Gleim et al. 2020) aimed at interlinking the multiple Digital Shadows in a holistic knowledge structure.

**Acknowledgments.** This work was supported in part by the Fraunhofer CCIT Research Cluster, and in part by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC-2023 Internet of Production – 390621612. I would like to thank numerous collaborators in these projects, especially Christoph Quix and István Koren (deputy area coordinators for the IoP infrastructure), the overall IDS project manager Boris Otto, and my IoP co-speakers Christian Brecher, Günter Schuh as well as IoP manager Matthias Brockmann.

**References**

Beutel, M.C., GöKay, S., Kluth, W., Krempels, K.-H., Samsel, C., Terwelp, C.: Product oriented integration of heterogeneous mobility services. In: IEEE 17th International Conference Intelligent Transportation Systems (ITSC), Qingdao, China (2014)

Cappiello, C., Gal, A., Jarke, M., Rehof, J.: Data ecosystems–sovereign data exchange among organizations. Dagstuhl Rep. 9(9), 66–134 (2019)
Dalibor, M., et al.: Model-driven development of a digital twin for injection molding. In: Proceedings CAiSE 2020, Grenoble, France (2020, to appear)

Eitel, A., et al.: Usage control in the industrial data space. Fraunhofer IESE Report 056.17/E (2017)

Gans, G., Jarke, M., Kethers, S., Lakemeyer, G.: Continuous requirements management for organization networks—a (dis-)trustful approach. Requirements Eng. 8(1), 4–22 (2003)

Gawer, A.: Bridging different perspectives on technological platforms—towards and integrative framework. Res. Policy 43(7), 1239–1249 (2014)

Gleim, L., Pennekamp, J., Liebenberg, M., et al.: FactDAG—formalizing data interoperability in an internet of production. IEEE Internet Things J. 4(7), 3243–3253 (2020)

Hai, R., Quix, C., Wang, D.: Relaxed functional dependency discovery in heterogeneous data lakes. In: Laender, A.H.F., Pernici, B., Lim, E.-P., de Oliveira, J.P.M. (eds.) ER 2019. LNCS, vol. 11788, pp. 225–239. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33223-5_19

Halevy, A., Franklin, M., Maier, D.: Principles of data space systems. In: Proceedings 25th ACM SIGMOD-PODS, pp. 1–9 (2006)

Jarke, M.: Data spaces: combining goal-driven and data-driven approaches in community decision and negotiation support. In: Schoop, M., Kilgour, D.M. (eds.) GDN 2017. LNBIIP, vol. 293, pp. 3–14. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63546-0_1

Jarke, M., Loucopoulos, P., Lyttilen, K., Mylopoulos, J., Robinson, W.: The brave new world of design requirements. Inf. Syst. 36(7), 992–1008 (2011)

Jarke, M., Quix, C.: On warehouses, lakes, and spaces—the changing role of conceptual modeling for data integration. In: Cabot, J., Gómez, C., Pastor, O., Sancho, M.-R., Teniente, E. (eds.) Conceptual Modeling Perspectives, 231-245. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-67271-7_16

Liebenberg, M., Jarke, M.: Information systems engineering with digital shadows—concept and case studies. In: Proceedings CAiSE 2020, Grenoble, France (2020, to appear)

Loucopoulos, P., Kavakli, E., Chechina, N.: Requirements engineering for cyber physical production systems. In: Giorgini, P., Weber, B. (eds.) CAiSE 2019. LNCS, vol. 11483, pp. 276–291. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21290-2_18

Otto, B., Jarke, M.: Designing a multi-sided data platform: findings from the international data spaces case. Electron. Markets 29(4), 561–580 (2019)

Otto, B., Steinbuß, S., Teuscher, A., Lohmann, S., et al.: Reference Architecture Model Version 3.0. International Data Spaces Association, Dortmund (2019)

Pfeiffer, A., Krempels, K.-H., Jarke, M.: Service-oriented business model framework—a service-dominant logic based approach for business modeling in the digital era. In: Proceedings 19th ICEIS Conference (Porto 2017), vol. 3, pp. 361–372 (2017)

Pennekamp, J., Glebke, R., Meisen, T., Quix, C., et al.: Towards an infrastructure enabling the internet of production. In: IEEE International Conference Industrial Cyber-Physical Systems (ICPS 2019), Taipei, TW, pp. 31–37 (2019)

Plato: Allegory of the Cave. The Republic VII (ca. 400 B.C.)

Pretschner, A., Hilty, M., Basin, D.: Distributed usage control. Commun. ACM 49(9), 39–44 (2005)

Tiwana, A., Konsynski, B.R., Bush, A.: Platform evolution—co-evolution of platform architecture, governance, and environmental dynamics. Inf. Syst. Res. 21(4), 675–687 (2010)

Van der Aalst, W.: Process Mining—Discovery, Conformance, and Enhancement of Business Processes. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19345-3

Yoo, Y., Henfridsson, O., Lyttilen, K.: The new organizing logic of digital innovation—an agenda for information systems research. Inf. Syst. Res. 21(4), 724–735 (2010)

Yu, E.: Modeling strategic relationships for process reengineering. In: Yu, E., et al. (eds.) Social Modeling for Requirements Engineering. Springer, Heidelberg (2011)