Data Platforms for Data Spaces

Amin Anjomshoaa, Santiago Cáceres Elvira, Christian Wolff, Juan Carlos Pérez Baún, Manos Karvounis, Marco Mellia, Spiros Athanasiou, Asterios Katsifodimos, Alexandra Garatzogianni, Andreas Trügler, Martin Serrano, Achille Zappa, Yury Glikman, Tuomo Tuikka, and Edward Curry

A. Anjomshoaa (✉)
Maynooth University, Maynooth, Ireland
e-mail: amin.anjomshoaa@mu.ie

S. C. Elvira
ITI – Instituto Tecnológico de Informática, Valencia, Spain

C. Wolff
ATB Bremen, Bremen, Germany

J. C. Pérez Baún
Atos Spain S.A., Madrid, Spain

M. Karvounis
Agroknow, Athens, Greece

M. Mellia
Politecnico di Torino, Torino, Italy

S. Athanasiou
Athena Research Center, Athens, Greece

A. Katsifodimos
TU Delft, Delft, Netherlands

A. Garatzogianni
Leibniz University Hannover, Hannover, Germany

A. Trügler
Know-Center GmbH, Graz, Austria

Graz University of Technology, Graz, Austria

M. Serrano · A. Zappa · E. Curry
Insight SFI Research Centre for Data Analytics, University of Galway, Galway, Ireland

Y. Glikman
Fraunhofer FOKUS, Berlin, Germany

T. Tuikka
VTT Technical Research Centre of Finland, Oulu, Finland

E. Curry et al. (eds.), *Data Spaces*, https://doi.org/10.1007/978-3-030-98636-0_3

Abstract In our societies, there is a growing demand for the production and use of more data. Data is reaching the point that is driving all the social and economic activities in every industry sector. Technology is not going to be a barrier anymore; however, where there is large deployment of technology, the production of data creates a growing demand for better data-driven services, and at the same time the benefits of the production of the data are at large an impulse for a global data economy, Data has become the business's most valuable asset. In order to achieve its full value and help data-driven organizations to gain competitive advantages, we need effective and reliable ecosystems that support the cross-border flow of data. To this end, data ecosystems are the key enablers of data sharing and reuse within or across organizations. Data ecosystems need to tackle the various fundamental challenges of data management, including technical and nontechnical aspects (e.g., legal and ethical concerns). This chapter explores the Big Data value ecosystems and provides a detailed overview of several data platform implementations as best-effort approaches for sharing and trading industrial and personal data. We also introduce several key enabling technologies for implementing data platforms. The chapter concludes with common challenges encountered by data platform projects and details best practices to address these challenges.

Keywords Data platforms · Data Spaces · Data ecosystem · Design

1 Introduction

Many industries and enterprises have recognized the real potential of Big Data value for exploring new opportunities and making disruptive changes to their business models. However, to realize the vision of Big Data value systems and create strong and sustaining Big Data ecosystems, several concerns and issues must be addressed. This includes [3] availability of high-quality data and data resources, availability of rightly skilled data experts, addressing legal issues, advancing technical aspects of data systems, developing and validating market-ready applications, developing appropriate business models, and addressing the societal aspects.

To foster, strengthen, and support the development and wide adoption of Big Data value technologies within an increasingly complex landscape requires an interdisciplinary approach that addresses the multiple elements of Big Data value. To this end, the introduction of the Big Data Value Reference Model (BDV-RM) [3] is an effort to address the common challenges and concerns of the Big Data value chain and create a data-driven ecosystem for Big Data. The BDVA Reference model is structured into core data processing concerns (horizontal) and cross-cutting concerns (vertical) as depicted in Fig. 1. The horizontal concerns include specific aspects along the data processing chain, starting with data collection and ingestion and extending to data visualization. On the other hand, vertical concerns address cross-cutting issues, which may affect all the horizontal concerns and involve nontechnical aspects.

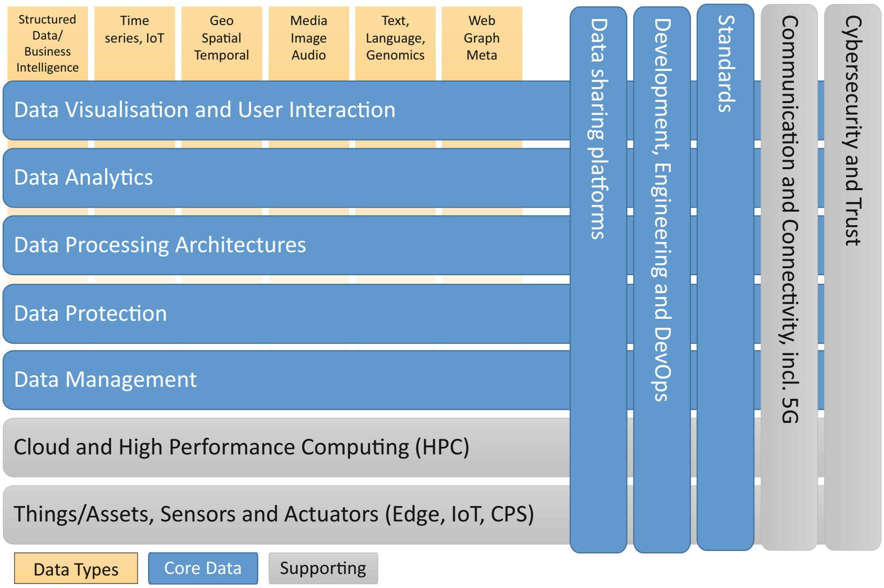
Structured Data/ Business Intelligence
Time series, IoT
Geo Spatial Temporal
Media Image Audio
Text, Language, Genomics
Web Graph Meta
Data Visualisation and User Interaction
Data Analytics
Data Processing Architectures
Data Protection
Data Management
Cloud and High Performance Computing (HPC)
Things/Assets, Sensors and Actuators (Edge, IoT, CPS)
Data sharing platforms
Development, Engineering and DevOps
Standards
Communication and Connectivity, incl. 5G
Cybersecurity and Trust
Data Types
Core Data
Supporting

Fig. 1 Big Data Value Reference Model

This book chapter first explores the Big Data value ecosystems. It introduces state-of-the-art data management systems that follow the BDV Reference Model to realize data value chains and data flows within ecosystems of intelligent systems. Then, we provide a detailed overview of several data platform implementations as best-effort approaches for sharing and trading industrial and personal data. We also compare the data management and Data Governance services of the data platform projects. Finally, the key enabling technologies for implementing data platforms will be introduced. We conclude this book chapter by providing an overview of common challenges encountered by data platform projects and best practices to address these challenges.

2 Big Data Value Ecosystems

A data ecosystem is a sociotechnical system that enables value to be extracted from data value chains supported by interacting organizations and individuals. Data value chains can be oriented to business and societal purposes within an ecosystem. The ecosystem can create the conditions for a marketplace competition between participants or enable collaboration among diverse, interconnected participants that depend on each other for their mutual benefit. Data ecosystems can be formed in different ways around an organization or community technology platforms or within or across sectors. This section introduces some best practices and proposed architectures to realize the Big Data value ecosystems.

2.1 Data Spaces and Data Platforms

The Big Data Value Association (BDVA)—that is, the private counterpart of the European Commission in the Big Data Value Public-Private-Partnership (BDV PPP)—defines data space as an umbrella term corresponding to any ecosystem of data models, datasets, ontologies, data sharing contracts, and specialized management services (i.e., as often provided by data centers, stores, repositories, individually, or within "data lake"'), together with soft competencies around it (i.e., governance, social interactions, business processes) [1]. These competencies follow a data engineering approach to optimize data storage and exchange mechanisms, preserving, generating, and sharing new knowledge.

In comparison, data platforms refer to architectures and repositories of interoperable hardware/software components, which follow a software engineering approach to enable the creation, transformation, evolution, curation, and exploitation of static and dynamic data in Data Spaces. To this end, a data platform would have to support continuous, coordinated data flows, seamlessly moving data among intelligent systems [2].

Although distinct, the evolution of the data space and data platform concepts goes hand in hand and needs to be jointly considered, and both can be considered the two faces of the same data economy coin. However, their complementary nature means that commercial solutions often do not distinguish between the two concepts. Furthermore, due to the particular requirements for the preservation of individual privacy, a distinction between technology and infrastructures that store and/or handle personal and other data has emerged. As a result, the evolution of industrial data platforms (considered key enablers of overall industrial digitization) and personal data platforms (services that use personal data, subject to privacy preservation, for value creation) has continued to follow different paths.

2.2 Gaia-X Ecosystem

Gaia-X[1] is a project to develop an efficient and competitive, secure, and trustworthy federation of data infrastructure and service providers for Europe, supported by representatives of business, science, and administration from European countries. Gaia-X follows the principles of openness and transparency of standards, interoperability, federation (i.e., decentralized distribution), and authenticity and trust.

The Gaia-X ecosystem is structured into a data ecosystem and the infrastructure ecosystem as depicted in Fig. 2. The data ecosystem enables Data Spaces as envisioned by the European data strategy, where data is exchanged, and advanced smart services are provided. The infrastructure ecosystem comprises building blocks from hardware nodes to application containers, where data is stored and services are executed, as well as networks for transmission of data between nodes and applications. addition, the infrastructure itself may be provided as a service.

[1] https://www.gaia-x.eu/

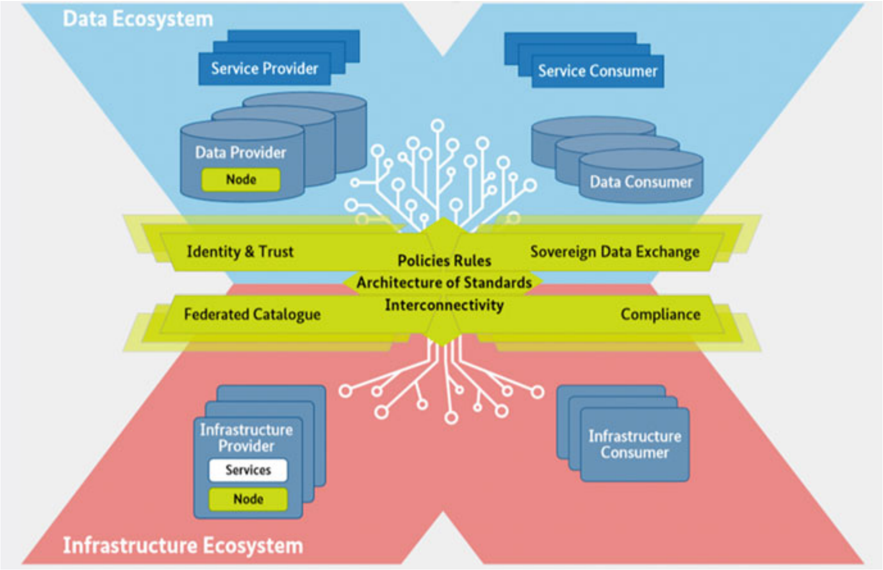

Fig. 2 Gaia-X architecture

3 Data Platform Project Portfolio

The data platform projects running under the umbrella of the Big Data Value Public-Private Partnership (BDV PPP) develop integrated technology solutions for data collection, sharing, integration, and exploitation to facilitate the creation of such a European data market and economy [3]. The portfolio of the Big Data value covers the data platform projects shown in Table 1. This table gives an overview of these projects, the type of data platform they develop, and the domain, the core enabling technologies, and the use cases they address. Each of these projects is briefly summarized in this section.

3.1 DataPorts Project

The DataPorts project[2] is devoted to creating a secure data platform that allows sharing the information between seaport agents in a reliable and trustworthy manner, with access permits and contracts to allow data sharing and the exploration of new Artificial Intelligence and cognitive services. It provides seaports with a secure and privacy-aware environment where the stakeholders can share data from different

[2] https://dataports-project.eu/

Table 1 Portfolio of the Big Data Value PPP covering data platforms

Project	Type	Technology	Use cases
DataPorts	Transportation	AI, blockchain, semantics	Seaport management
TheFSM	Food	AI, blockchain, semantics	Food supply chain
i3-Market	Generic data market support tools	Semantics, blockchain, OpenID	Automotive Manufacturing Wellbeing
OpertusMundi	Generic geodata market	Microservices, BPMN workflows	Geospatial data market
Trusts	Personal/industrial data market	Data encryption Blockchain KAN-based open data repositories Semantics	Data market Finance Telecom
smashHit	Personal/industrial data market	Semantics	Insurance, automotive industry, insurance, smart city
PimCity	Personal data market	Machine learning, data provenance, privacy-preserving	Generic data market
Kraken	Personal data market	Blockchain, privacy-preserving, self-sovereign identity, data encryption	Education health
DataVaults	Personal data market	Machine learning, blockchain	Sports Mobility Healthcare Tourism Smart home Smart city

sources to get real value, providing a set of novel AI and cognitive tools to the port community.

The platform takes advantage of huge data provided by stakeholders for improving existing processes and enabling new business models. To this end, the project offers several common analytics services such as auto model training and machine learning pipelines that seaports agents can reuse.

The Data Governance components of the project benefit from Semantic Web technologies to enable interoperability between stakeholders and blockchain technology that realizes the business rules via smart contracts. Figure 3 provides an overview of data services offered by DataPorts data platform.

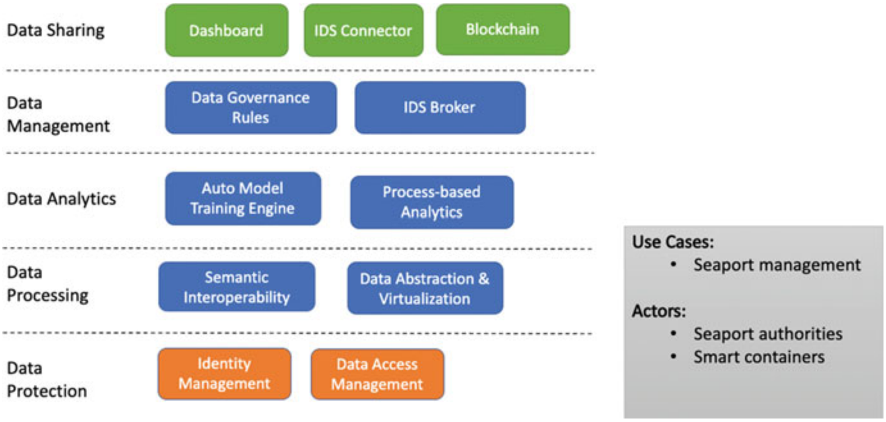

Fig. 3 System overview and data services of DataPorts data platform

3.2 TheFSM Project

TheFSM platform[3] aspires to ensure transparent and safe food production by digitizing food certification processes that assess safety via audits. More specifically, during the past 5 years, we have witnessed major changes in the food sector, with tremendous emphasis being put on food safety. A series of food safety scandals and health incidents have led to the international alignment of food safety standards through the Global Food Safety Initiative (GFSI). Governments also apply stricter policies and legislation, such as the integrated food safety policy of the European Commission and the US Food Safety Modernization Act (FSMA). There is increased pressure for the agri-food and grocery sector to ensure that their suppliers comply with food safety standards recognized by the GFSI. This translates into more pressure for all stakeholders in the supply chain to exchange data critical to food safety assessment and assurance in a timely, trusted, and secure manner. Finally, the global COVID-19 pandemic has further emphasized the need for supporting digital and remote auditing and certification processes.

The Food Safety Market (TheFSM) aims to deliver an industrial data platform that will significantly boost food certification in Europe. To achieve this goal, and as the food certification market is multifaceted, there is the need for all the actors in the food supply chain to share food safety data in a well-defined and automated way. Therefore, the platform aims to establish remote auditing in the European food market and serves as a data marketplace that enables all actors in the food chain to monitor, trace, and predict food safety risks in the food supply chain, to allow food safety inspectors and auditors to manage inspection/certification workflow

[3] https://foodsafetymarket.eu/

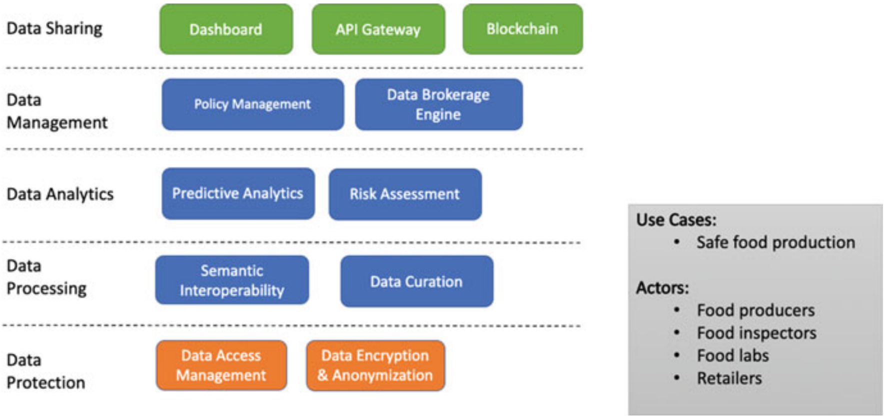

Fig. 4 System overview and data services of TheFSM data platform

digitally, and to allow farmers and food producers to manage their resources and their certification data.

The platform provides data curation and semantic enrichment services to create and manage a Knowledge Graph of domain objects. Furthermore, the platform benefits from blockchain technology to provide a collaborative hub for connecting organizations aiming to work together and solve complex supply chain challenges.

Eventually, TheFSM aspires to catalyze the digital evolution of global food certification's traditional but very data-intensive business ecosystem. Figure 4 provides an overview of data services offered by TheFSM data platform.

3.3 i3-MARKET Project

It has been largely discussed that there is a growing demand for a global data economy, where the different data stakeholders can participate in the distribution of the benefits from selling/trading data assets. The i3-MARKET project[4] addresses this growing demand from the perspective of a single European Data Market Economy by innovating marketplace platforms, enabling them with software artifacts that allow the deployment of data-related services, and demonstrating that data economy growth is possible with industrial implementations. The i3-MARKET solution(s) aims at providing technologies for trustworthy (secure and reliable), data-driven collaboration and federation of existing and new future marketplace platforms, with particular attention on industrial data. Furthermore, the i3-MARKET architecture is

[4] https://www.i3-market.eu/

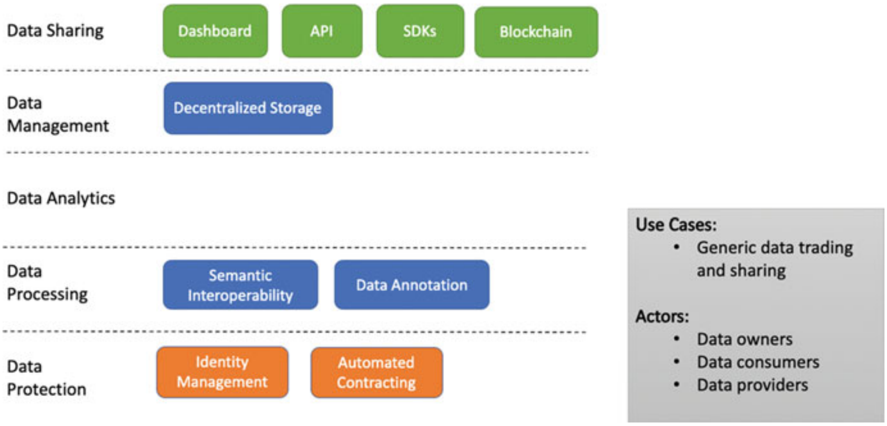

Fig. 5 System overview and data services of i3-Market data platform

designed to enable secure and privacy-preserving data sharing across Data Spaces and marketplaces by deploying a backplane across operational data marketplaces.

The i3-MARKET project does not try to create another new marketplace, involving the multiple data marketplace characteristics and functionalities; rather, it implements a backplane solution that other data marketplaces and Data Spaces can use to expand their market data offering capacities; facilitate the registration and discovery of data assets; facilitate the trading and sharing of data assets among providers, consumers, and owners; and provide tools to add functionalities they lack for a better data sharing and trading processes across domains. By bringing together data providers (supply side) and data consumers (demand side), i3-MARKET acts as an enabler for data monetization, realizing promising business ideas based on data trading, and trustworthy and data-driven collaboration. This way, i3-MARKET is the missing link acting as reference implementation that will allow all the interested parties to connect while offering incentives to data owners, data providers, and data consumers to engage in data trading. It may also serve as best practices for enabling data-driven economy and pave the way to the European data sharing economy in a safer, secured, and fair manner. The i3-MARKET project targets the possibility to interconnect data assets in a distributed manner enabling a federated query system that facilitates increasing the data offerings without the need to collect and host data locally. In this form any data marketplace that registers to the i3-MARKET ecosystem is able to provide access to cross-domain description and use the smart contract approach to be able to allow the access to the data asset remotely. Figure 5 provides an overview of data services offered by i3-Market data platform.

3.4 OpertusMundi

The OpertusMundi project[5] aims to deliver a trusted, secure, and scalable pan-European industrial geospatial data market, Topio,[6] acting as a single point for the streamlined and trusted discovery, sharing, trading, remuneration, and use of geospatial data assets, guaranteeing low cost and flexibility to accommodate current and emerging needs of data economy stakeholders regardless of size, domain, and expertise.

Topio empowers *geospatial data suppliers* to trade their assets under (a) *homogenized*, *configurable, digital*, and *automated* contracting facilities enforceable across EU; (b) multiple *standardized pricing models* and *tiers* suited to the type of their assets and business models; (c) full autonomy in *publishing*, *vetting*, and *monitoring* the sales and use of their assets via rich integrated analytics and IPR protection schemes; (d) novel *monetization schemes* by *automatically* exposing *data* as *services* created and operationalized by Topio in *a revenue-sharing* scheme; and (e) unrestricted *opt-in/out* of its services. From the *consumer's* perspective, Topio enables them to *fairly* purchase assets that are *fit for purpose* via (a) rich automated metadata for traded assets (i.e., *data profiling*) independently provided by Topio to support *informed purchasing decisions*, (b) clear and transparent *terms, conditions*, and *pricing* for assets *before purchase*, (c) automated digital *contracting* and *payments* with dispute resolution provided by Topio, and (d) *streamlined*, *low-effort,* and *direct use* of purchased and private assets via a plethora of web services. For all types of users, including those not actively trading or purchasing data, Topio enables them to *use* and *extract value* from geospatial data through a plethora of low-cost and intuitive *value-added services*, ranging from cloud storage and custom maps to Jupyter notebooks and analysis bundles for select thematic domains.

Topio, as a sustainable commercial endeavor and business entity, is designed and built on the principles of *trust*, *fairness,* and adherence to *law*. On a technical level, it is *fully auditable* via automated BPMN workflows, with all transactions taking place by KYB/KYC-validated entities under anti-money laundering (AML) safeguards. On a legal level, it is in full conformance with the EU's current and emerging legal framework on Data Spaces and markets, e-marketplaces, consumer rights, competition, and especially the Data Governance Act.[7] Finally, on a business level, Topio's business model is founded on our vision to *grow* and *serve* the emerging data economy in the EU, estimated[8] to reach 550b€ in size by 2025,[9] with 99b€ in data supplier revenues. Considering that ~80% of data are anecdotally considered as *geospatial*, and with 583K EU data users and 173K suppliers by 2025, Topio's

[5] https://www.opertusmundi.eu/

[6] https://topio.market/

[7] Data Governance Act, COM/2020/767 final.

[8] https://ec.europa.eu/digital-single-market/en/news/european-data-market-study-update

[9] 829b€ according to https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy

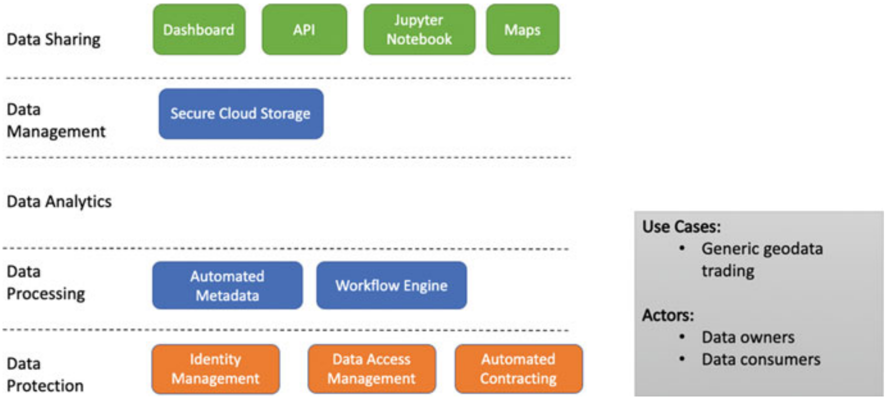

Fig. 6 System overview and data services of OpertusMundi data platform

impact in materializing this vision can be substantial. For this reason, Topio's business model and service offerings *do not, and will not, impose fees* on data trading. Instead, Topio generates its profits *solely* from operationalizing supplier-provided *data as services* and *subscriptions* from its value-added services. Figure 6 provides an overview of data services offered by OpertusMundi data platform.

3.5 TRUSTS Project

The TRUSTS project[10] aims to ensure the sustainable business uptake of secure data markets by enabling a fully operational and GDPR-compliant European data marketplace for personal and industrial data in the finance and telecom sectors, while allowing the integration and adoption of future platforms. To this end, the platform provides services to identify and overcome legal, ethical, and technical challenges of cross-border data markets. The platform follows the reference architecture designed by the International Data Spaces (IDS) Association which uses Semantic Web technologies for describing data schemas to configure connectors and interpret the data shared through these connectors. Furthermore, the proposed approach aims to create trust between participants through certified security functions, to allow secure collaboration over private data, and to establish governance rules for data usage and data flows. The IDS architecture ensures data sovereignty for those who make data available in data ecosystems. Figure 7 provides an overview of data services offered by TRUSTS data platform.

[10] https://www.trusts-data.eu/

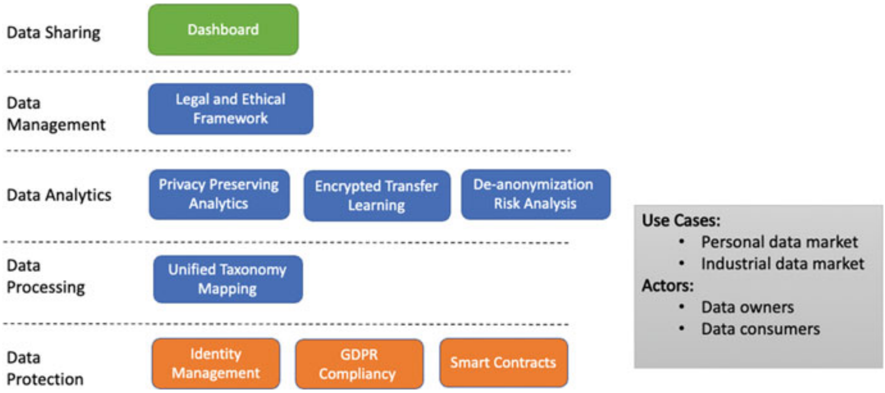

Fig. 7 System overview and data services of TRUSTS data platform

3.6 smashHit Project

The objective of smashHit[11] is to assure trusted and secure sharing of data streams from both personal and industrial platforms, needed to build sectorial and cross-sectorial services. The project establishes a framework for processing data owner consent and legal rules and effective contracting, as well as joint security and privacy-preserving mechanisms. The vision of smashHit is to overcome obstacles in the rapidly growing data economy, which is characterized by heterogeneous technical designs and proprietary implementations, lacking business opportunities due to the inconsistent consent and legal rules among different data sharing platforms, actors, and operators. By using the smashHit project solutions, it is expected to achieve improved citizen trust (by providing data owners awareness of their given consent), improved OEM and data customer trust (due to fingerprinted data to ensure traceability/unchangeability along the value chain as well as due to consent tracing), simplified consent process (by providing certification of consent and a single point of consent management), and support in consent/contract generation (facilitating the generation of legally binding contracts, taking into account relevant legislation/legal rules). Figure 8 provides an overview of data services offered by smashHit data platform.

3.7 PimCity Project

The PimCity project[12] aims to increase transparency in online data markets by giving users control over their data, ensuring that citizens, companies, and

[11] https://smashhit.eu/

[12] https://www.pimcity-h2020.eu

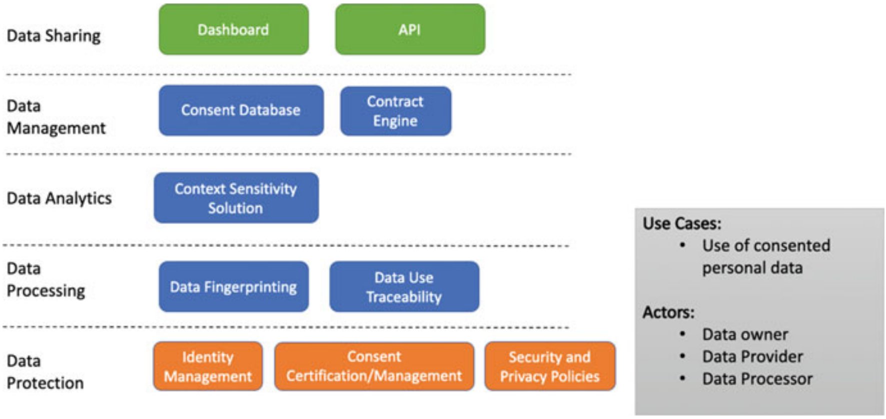

Fig. 8 System overview and data services of smashHit data platform

organizations are informed and can make respectful and ethical use of personal data. The project follows a human-centric paradigm aimed at a fair, sustainable, and prosperous Digital Society. The sharing of personal data is based on trust and a balanced and fair relationship between individuals, businesses, and organizations.

The project provides a PIMS Development Kit (PDK) that allows developers to engineer and experiment with new solutions. It allows them to integrate new data sources and connect them to new services. The PDK focuses on interoperability, which is at the same time the most significant challenge because it requires a process of standardization of consent mechanisms, formats, and semantics. All platform components offer Web APIs that are documented using the Open API specifications to allow seamless integration. This enables communications and interactions among components in the PDK, easing integration with existing PIMS, and the design and development of new ones. Figure 9 provides an overview of data services offered by PimCity data platform.

3.8 *KRAKEN Project*

KRAKEN[13] (brokerage and market platform for personal data) aims to develop a trusted and secure personal data platform with state-of-the-art privacy-aware analytics methods that return the control of personal data to users. The project also aims to enable the sharing, brokerage, and trading of potentially sensitive personal data by returning the control of this data to citizens throughout the entire data lifecycle. The project combines, interoperates, and extends the best results from two

[13] https://www.krakenh2020.eu/

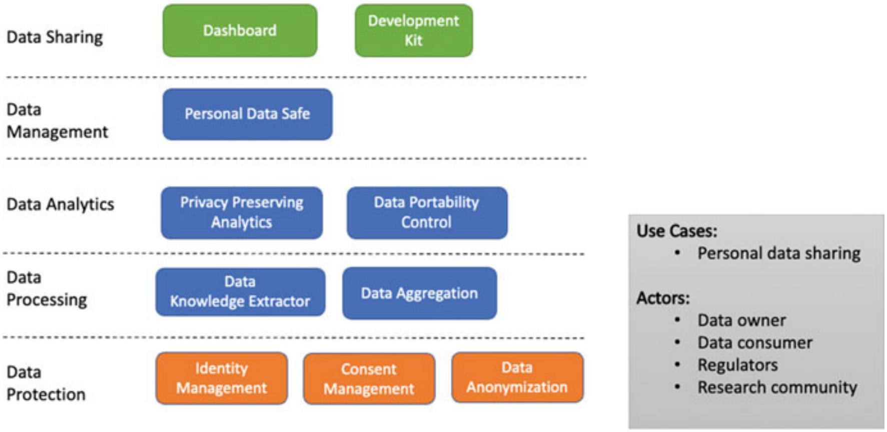

Fig. 9 System overview and data services of PimCity data platform

existing mature computing platforms developed within two H2020 actions, namely, CREDENTIAL[14] and MyHealthMyData.[15]

The project addresses the challenge of removing obstacles that prevent citizens from controlling and widely sharing their personal data. With this objective, KRAKEN is investigating data processing mechanisms working in the encrypted domain to increase security, privacy, functionality, and scalability for boosting trust. In this sense, KRAKEN will provide a highly trusted, secure, scalable, and efficient personal data sharing and analysis platform adopting cutting-edge technologies. The KRAKEN project is based on three main pillars:

- The self-sovereign identity (SSI) paradigm provides user-centric access control to data. The data owner controls their data by using an SSI mobile app where the verifiable credentials and key material are stored.
- The cryptographic techniques support the other two pillars, such as functional encryption (FE) and secure multi-party computation (SMPC). These tools enable building a data-analytics-as-a-service platform integrated with the marketplace. They also ensure end-to-end secure data sharing in terms of confidentiality and authenticity.
- Data marketplace brings together the other two pillars, demonstrating in two high-impact pilots health and education the applicability of the KRAKEN solution. The marketplace acts as an open and decentralized exchange system connecting data providers and data consumers, leveraging a blockchain network facilitating the business and legal logic related to the data transactions.

[14] https://credential.eu/

[15] http://www.myhealthmydata.eu/

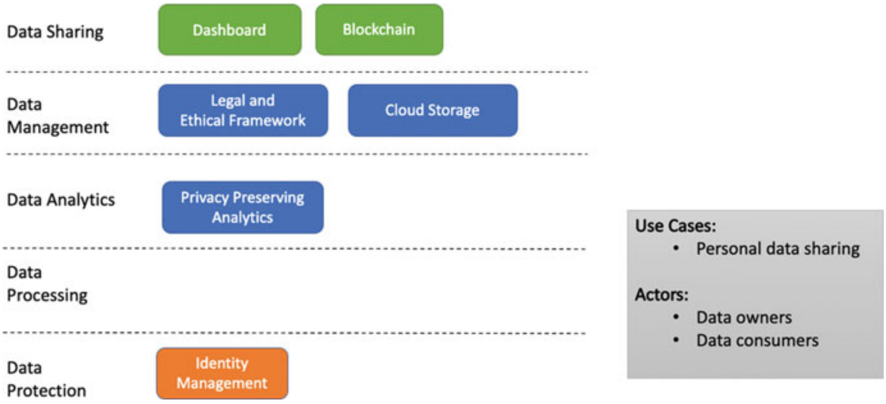

Fig. 10 System overview and data services of Kraken data platform

To follow regulatory frameworks and be GDPR and eIDAS compliant, an ethical and legal framework is implemented, affecting both the design and the implementation aspects. Figure 10 provides an overview of data services offered by Kraken data platform.

3.9 DataVaults Project

DataVaults[16] aims to deliver a framework and a platform that has personal data, coming from diverse sources in its center and that defines secure, trusted, and privacy-preserving mechanisms allowing individuals to take ownership and control of their data and share them at will, through flexible data sharing and fair compensation schemes with other entities (companies or not). Furthermore, data queries and analysis in DataVaults will allow the linking and merging of data from various sources and combining those with personal data, based on the DataVaults core data model. These activities, which rely on the semantic annotation of data and the curation of those to make them linkable, will raise the economic value of both personal and other kinds of data, as more detailed and interesting insights will be generated.

The overall approach will rejuvenate the personal data value chain, which could from now on be seen as a multi-sided and multi-tier ecosystem governed and regulated by smart contracts which safeguard personal data ownership, privacy, and usage and attributes value to the ones who produce it. Addressing the concerns on privacy, ethics and IPR ownership over the DataVaults value chain is one of the cornerstones of the project. It aims to set, sustain, and mobilize an ever-growing

[16] https://www.datavaults.eu/

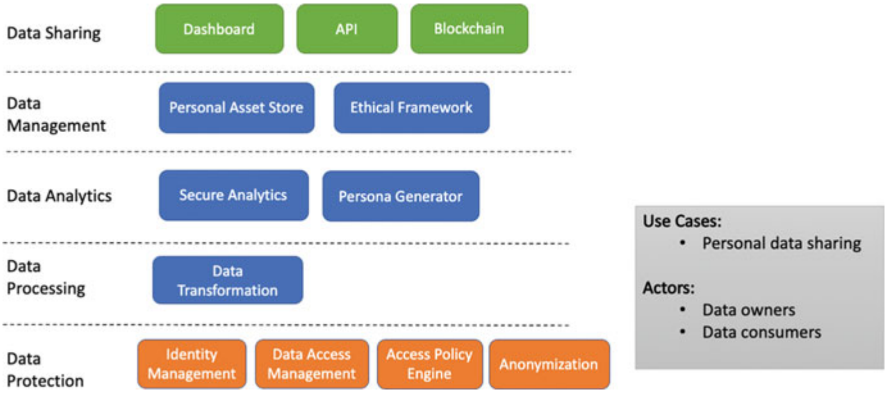

Fig. 11 System overview and data services of DataVaults data platform

ecosystem for personal data and insights sharing and for enhanced collaboration between stakeholders (data owners and data seekers) on the basis of DataVaults personal data platform's extra functionalities and methods for retaining data ownership, safeguarding security and privacy, notifying individuals of their risk exposure, as well as on securing value flow based on smart contract. Figure 11 provides an overview of data services offered by DataVaults data platform.

4 Comparison of Data Management Services

The projects presented in the previous section will be explored based on their data management and Data Governance features. The comparison is based on the following primary requirements for a catalogue and entity management service (EMS) [4], which are needed to support the incremental data management approach of Data Spaces:

- Data Source Registry and Metadata: The requirement to provide a registry for static and dynamic data sources and their descriptions.
- Entity Registry and Metadata: The requirement to provide a registry for entities and their descriptions.
- Machine-Readable Metadata: The requirement to store and provide metadata about data sources and entities in machine-readable formats using open standards such as JavaScript Object Notation (JSON) and Resource Description Framework (RDF).
- Schema Mappings: The capability to define mappings between schema elements.
- Entity Mappings: The capability to define mappings between entities.
- Semantic Linkage: The capability to define semantic relationships and linkages among schema elements and entities.

- Search and Browse Interface: The requirement to provide a user interface over the catalogue and EMS, which allows searching and browsing all the stored elements.
- Authentication and Authorization: The requirement to verify the credentials of users and applications accessing the catalogue and EMS, which can limit access to sources/entities based on access policies or rules.
- Data Protection and Licensing: The requirement to fulfill the privacy and confidentiality requirements of data owners and provide licensing information on the use of data.
- Provenance Tracking: The requirement of tracking the lineage of changes made to the catalogue and EMS by users and applications.

	DP	FSM	I3M	OM	T	SH	PC	K	DV
Data Source Registry and Metadata	+	+	+	+	+	+	+	+	+
Entity Registry and Metadata	+	+	+	+	+	+	+	+	+
Machine-Readable Metadata	+	+	+	+	+		+	+	+
Schema Mappings			+	+	+		+		+
Entity Mappings		+			+				+
Semantic Linkage	+	+	+	+	+				+
Search and Browse Interface	+	+	+	+	+	+	+	+	+
Authentication and Authorization	+	+	+	+	+	+	+	+	+
Data Protection and Licensing		+		+	+	+	+	+	+
Provenance Tracking	+	+	+	+			+	+	

5 Key Enabling Technologies

Data platform implementations address a set of common and known problems and challenges which can be tackled through existing guidelines, appropriate technologies, and best practices. In this section we provide an overview of some key enabling technologies that have broad application in the implementation of data platforms.

5.1 Semantics

Semantics is becoming increasingly important within Data Spaces and potentially becomes an area of competitive advantage in European data space and data market projects where semantics plays the role of federator and ecosystem facilitator. In order to facilitate the use of data platforms, they should be able to semantically annotate and enhance the data without imposing extra effort on data owners and data producers. The semantically enhanced data will improve various data processes and unlock data silos using interoperability standards and efficient technologies of the

Semantic Web domain. For instance, data integration, one of the hardest challenges in computer science, benefits significantly from semantics. The challenging issue in data integration is that people have different conceptualizations of the world. As a result, we would need a computational semantics layer to automate the process of data integration. The realization of the semantic layer could range from implementing taxonomies and metadata schema to more profound knowledge engineering approaches, including ontologies and semantic deduction. Semantic technologies and related solutions are also used for indexing and discovery of data services and creating service repositories in the context of data platform ecosystems.

5.2 Blockchain and Smart Contracts

As data and its corresponding services move beyond the organization's borders, we need mechanisms that support data integrity and traceability. In decentralized data platforms where peers need to share and consume data resources, trustworthiness and transparency of data processes are of great importance. Such platforms should provide features such as data provenance, workflow transparency, and authenticity of data providers and data consumers. To this end, blockchain offers disruptive methods to view the Data Governance, open data, and data ownership problems from a new perspective.

Recent advances in blockchain technology have shown the potential of building trustworthy data sharing solutions while maintaining a sufficient level of transparency in decentralized ecosystems. To this end, several data platforms have employed smart contracts as the conceptual basis for capturing and realization of business requirements. Furthermore, the smart contract and blockchain-driven approaches also incentivize user engagement and monetizes data platform solutions.

5.3 AI and Machine Learning

The success of AI and machine learning approaches is determined by the availability of high-quality data, which allows running various statistical learning methods to build efficient models and processes. To this end, data platforms are one of the key enablers to AI and machine learning processes. The value-added services of data platforms make high-quality and trustworthy data available to data consumers who use the data to create innovative solutions and turn data into products and services. Furthermore, data platforms apply machine learning and AI processes to improve data quality and enrich the data with the required information during the data preparation pipelines. Due to the increasing amount of data and the growing need for data and data services, applying AI and machine learning methods in data processes is inevitable. Several data platforms are already equipped with such processes, which range from data analysis (e.g., data cleaning and data integration) and data enrichment (e.g., data annotation and clustering) to more advanced processes such as natural language processing (NLP) and automatic enforcement of legal and ethical requirements (e.g., identifying sensitive data and data anonymization).

6 Common Challenges and Lessons Learned

The design and implementation of data platforms pose significant technical challenges such as data integration, interoperability, privacy-related issues, and non-technical challenges such as legal and technical issues, engaging end-users, and organizational challenges. This section provides an overview of such challenges encountered by data platform projects and discusses some best practices to address these challenges.

6.1 AI and Machine Learning Challenges

The recent advances in AI and machine learning have led to widespread use in data-driven use cases and industries [5]. Although the results are promising, and in many cases comparable to human performance, there are several situations where the outcomes are unreliable and require human intervention to address irregular and unpredictable situations. For instance, the lack of comprehensive datasets in some cases may lead to algorithmic discrimination against minorities or misleading outcomes.

In the AI and machine learning domains, high-quality data plays a significant role and greatly determines the quality of outcomes. As such, data platforms need to integrate best practices for data quality into their architecture in order to provide trustworthy and reliable data sources.

One particular challenge ahead is the automation of AI methods in data platforms, which depends on several configurations such as choice of algorithm, configuring relevant parameters of selected algorithms, and identification of features from available datasets. So, including metadata and machine-readable description of AI and machine learning components and including them in transfer learning processes seem to be the key to addressing these issues.

Another challenge in machine learning services of data platforms is to supplement the outcomes with human-understandable explanations. Explainable AI is an active field of research in the AI domain, and the data platform community needs to consider the explainability feature in data platform processes when appropriate.

6.2 Legal Challenges

Data platforms require to implement and realize legal requirements. In the case of domain-specific data platforms that deal with known vendors and predetermined business requirements, the policy and regulation are assessed and included in the architecture of the corresponding data platform. However, in the case of generic data platforms such as data markets, the uncertainty of policy and regulations makes this task very difficult, if not impossible. The legal challenges span many policies and regulations, including privacy and data protection law, national/international data protection legislation (e.g., GDPR), human rights law (e.g., EU Convention on Human Rights), and business regulations.

In addition to identifying the relevant regulations for different datasets and use cases, the regulations need to be captured and articulated in a machine-readable way and be integrated into data platform processes.

6.3 Ethical Challenges

Moral and ethical guidelines are challenging parts of data platform implementation. Similar to legal challenges, there is a handful of relevant policies and guidelines for justified use of data in data platforms. However, there is no one-size-fits-all scenario, and we need to define the ethical regulations based on the specific requirements of the target domain.

The ethical challenges are weighted more importantly when dealing with the storage and sharing of personal information. Therefore, we would need privacy-aware analytics methods to realize the moral and ethical requirements. Furthermore, in the case of personal information, we would also need fine-grained consent management processes that clearly define and include user preferences and the conditions of granting and revoking data usage permissions [6].

6.4 Sustainability Challenges

In data platforms, sustainability includes two dimensions: sustainability of software architecture and sustainability of data services. The development of sustainable architecture is a well-studied domain. As a result, there are a handful of guidelines and solutions for implementing a sustainable software architecture that can be applied to the specific case of data platform systems. In contrast, the sustainability of data services in generic data platforms is not a straightforward task. The reason is that data services are usually built on top of a complex web of data, policies, and regulations that might change over time. For instance, if the international e-privacy regulation is changed, data platforms need to evaluate the new requirements and make sure the data platform complies with new requirements, which is a challenging task. Also, the sustainability of data services is even more complicated if the policy changes need to be applied to legacy data processes and existing data processes. For instance, if the data sharing process uses blockchain technology, reconsidering shared data might be infeasible due to the immutable nature of public blockchains.

6.5 User Engagement Challenges

Active participation of users from the very early stages of a data platform project is a fundamental element for the successful implementation and realization of data platforms. User participation guarantees continued support and sustainable development of data platform systems in the future. Therefore, adopting a user-oriented approach, while analyzing business requirements and exploring legal and ethical aspects through user stories, is a key factor here. However, despite the fundamental role of user engagement in data platform projects, a handful of challenges make user involvement difficult. For instance, lack of trust for sharing personal and industrial data, before realizing a data platform and envisioning benefits and gained values for each stakeholder, is one of the main challenges.

7 Conclusion

Big Data ecosystems offer enormous potential to support cross-organization data sharing and trading in a trusted, secure, and transparent manner. This chapter presented several data platform projects and highlighted their common features and services. These platforms provide various services and facilitate reliable and transparent data flow between systems. However, to achieve maximum economic and societal benefits, several challenges in AI and machine learning, and ethical, legal, and sustainability domains need further investigation.

References

1. Big Data Value Association. (2019). *Towards a European Data Sharing Space: Enabling data exchange and unlocking AI potential*. BDVA. http://www.bdva.eu/node/1277
2. Curry, E., & Ojo, A. (2020). Enabling knowledge flows in an intelligent systems data ecosystem. In *Real-time linked dataspaces* (pp. 15–43). Springer.
3. Zillner, S., Curry, E., Metzger, A., & Auer, S. (2017). European big data value partnership strategic research and innovation. *Agenda, 2017*.
4. ul Hassan, U., Ojo, A., & Curry, E. (2020). Catalog and entity management service for internet of things-based smart environments. In *Real-time Linked Dataspaces* (pp. 89–103). Springer.

5. Zillner, S., Bisset, D., Milano, M., Curry, E., Södergård, C., & Tuikka, T. (2020). *Strategic research, innovation and deployment agenda: AI, data and robotics partnership*.

6. EC. (2020). *Communication: A European strategy for data*. Retrieved from https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf