A Framework for Big Data Sovereignty: The European Industrial Data Space (EIDS)

Christoph Mertens, Jesús Alonso, Oscar Lázaro, Charaka Palansuriya, Gernot Böge, Alexandros Nizamis, Vaia Rousopoulou, Dimosthenis Ioannidis, Dimitrios Tzovaras, Rizkallah Touma, Miquel Tarzán, Diego Mallada, Paulo Figueiras, Ruben Costa, Diogo Graça, Gisela Garcia, Begoña Laibarra, Aitor Celaya, Piotr Sobonski, Azzam Naeem, Alberto Mozo, Stanislav Vakaruk, J. Enrique Sierra-García, Antonio Pastor, Juan Rodriguez, Marlène Hildebrand, Tomasz Luniewski, Wojciech Zietak, Christoph Lange, Konstantinos Sipsas, and Athanasios Poulakidas

C. Mertens
International Data Spaces e.V. (IDSA), Berlin, Germany
e-mail: christoph.mertens@internationaldataspaces.org

J. Alonso (✉) · O. Lázaro
Asociación de Empresas Tecnológicas Innovalia, Bilbao, Spain
e-mail: jalonso@innovalia.org; olazaro@innovalia.org

C. Palansuriya
EPCC, The University of Edinburgh, Edinburgh, UK
e-mail: charaka@epcc.ed.ac.uk

G. Böge
FIWARE Foundation e.V, Berlin, Germany
e-mail: gernot.boege@fiware.org

A. Nizamis · V. Rousopoulou · D. Ioannidis · D. Tzovaras
Centre of Research and Technology Hellas (CERTH), Thessaloniki, Greece
e-mail: alnizami@iti.gr; vrousop@iti.gr; djoannid@iti.gr; Dimitrios.Tzovaras@iti.gr

R. Touma · M. Tarzán
Fundació i2CAT, Barcelona, Spain
e-mail: rizkallah.touma@i2cat.net; miquel.tarzan@i2cat.net

D. Mallada
GESTAMP, Pillaipakkam, Tamil Nadu, India
e-mail: dmallada@gestamp.com

P. Figueiras · R. Costa
UNINOVA, Guimarães, Portugal
e-mail: paf@uninova.pt; rddc@uninova.pt

D. Graça · G. Garcia
Volkswagen Autoeuropa (VWAE), Lisbon, Portugal
e-mail: Diogo.Graca@volkswagen.pt; Gisela.Garcia@volkswagen.pt

Abstract The path that the European Commission foresees to leverage data in the best possible way for the sake of European citizens and the digital single market clearly addresses the need for a European Data Space. This data space must follow the rules, derived from European values. The European Data Strategy rests on four pillars: (1) Governance framework for access and use; (2) Investments in Europe's data capabilities and infrastructures; (3) Competences and skills of individuals and SMEs; (4) Common European Data Spaces in nine strategic areas such as industrial manufacturing, mobility, health, and energy. The project BOOST 4.0 developed a prototype for the industrial manufacturing sector, called European Industrial Data Space (EIDS), an endeavour of 53 companies. The publication will show the developed architectural pattern as well as the developed components and introduce the required infrastructure that was developed for the EIDS. Additionally, the population of such a data space with Big Data enabled services and platforms is described and will be enriched with the perspective of the pilots that have been build based on EIDS.

B. Laibarra · A. Celaya
Software Quality Systems (SQS), Oswego, NY, USA
e-mail: blaibarra@sqs.es; acelaya@sqs.es

P. Sobonski · A. Naeem
United Technology Research Centre Ireland, Cork, Ireland
e-mail: piotr.sobonski@rtx.com; azzam.naeem@rtx.com

A. Mozo · S. Vakaruk
Universidad Politécnica de Madrid (UPM), Madrid, Spain
e-mail: a.mozo@upm.es; stanislav.vakaruk@upm.es

J. E. Sierra-García
Universidad de Burgos, JRU ASTI-UBU, Burgos, Spain
e-mail: jesierra@ubu.es; enriquesg.cext@asti.tech

A. Pastor · J. Rodriguez
Telefonica I+D, Madrid, Spain
e-mail: antonio.pastorperales@telefonica.com; juan.rodriguezmartinez@telefonica.com

M. Hildebrand
EPFL, Lausanne, Switzerland
e-mail: marlene.hildebrand@epfl.ch

T. Luniewski · W. Zietak
CAPVIDIA, Houston, TX, United States
e-mail: tl@capvidia.com; wz@capvidia.com

C. Lange
Fraunhofer Institute for Applied Information Technology FIT, RWTH Aachen University, Aachen, Germany
e-mail: christoph.langebever@fit.fraunhofer.de

K. Sipsas · A. Poulakidas
INTRASOFT, Mumbai, Maharashtra, India
e-mail: Konstantinos.SIPSAS@intrasoftintl.com; Athanasios.Poulakidas@intrasoft-intl.com

1 Introduction

Over the last two decades, data economy has emerged as a global growing trend, with heterogeneity of data sources becoming available. Economic development has evolved from a public open data or science data smart usage to the mass adoption and exploration of the value of industrial Big Data. And now, a new paradigm shift looms, evolving toward the economy of data sharing or common data.

Industry is now starting to learn how to extract value out of its own industrial data to gain industrial competitiveness; companies are also starting to realize that it is extremely unlikely that a single platform, let alone a single company, will drive the industrial data business. Hence, the development of new services and products relying on access to common data in the industrial environment calls for the transformation of various aspects. One of the central challenges in the production environment is the handling of the data that is generated, whether it comes from sensors of machines, in planning processes, or in product development. The news that this data is very valuable has now reached the broad base of society. The consequence, however, is not that the data is purposefully turned into money, but that this data treasure is locked away in many cases.

With the Data Governance Act and AI regulation, Europe is already laying the foundations for a European common data space. However, this policy framework needs to be matched with transformation at the industrial level of data sharing business culture and convergence of data infrastructures and soft data space services. It must be acknowledged that, on one hand, the common data explosion will arrive when industry exhausts its ability to extract value out of its own industrial data. On the other hand, it must also be acknowledged that it is very unlikely that the B2B common data economy will be based on access to the raw data entities themselves.

On the contrary, data ecosystems and future value chains will more likely be developed on sovereign access to well-curated and well-controlled data endpoints of high quality. Hence, the most urgent need is to develop not only a data culture but a data sharing or common data one.

Therefore, the development of data ecosystems should be envisioned as an evolutionary rather than a revolutionary process that will support new business models and disruptive value chains. This might be a chicken-egg problem. Data sharing and the data economy bring challenges that must be faced, like trust and openness, interoperability and cost, continuity, and controlled free flow of data. These are challenges, yes, but what is most important is that all of these lead to necessities.

In order to avoid a data divide and provide truly free access to a European common sovereign data space, there's a need for both a comprehensive digital

infrastructure, not only a physical one but a soft one, related to all the services for sovereign data sharing and controlled data usage among different organizations. This should be accompanied by specific industrial agreements that will define the governance rules of future value chains. A well-balanced development of such dimensions is the best guarantee for the growth of the emerging and embryonic Data Spaces that the European industry is already putting in place to keep the leadership of the industrial (B2B) data and intelligence economy. The industry is already collaborating in finding an agreement on the design principles that should drive the development of such Data Spaces and ecosystems. One of the first large-scale implementations of an industrial embryonic data space has been implemented in the Boost 4.0 project [1]: the European Industrial Data Space (EIDS).

The EIDS is based on the International Data Spaces (IDS) Reference Architecture Model developed by the International Data Spaces Association (IDSA), a nonprofit association with more than 140 members from all over the world. The association aims to give back full control over data to companies, by coming up with the so-called IDS standard.

The main goal of the IDS standard is to provide trust between the different participants in the data space. To achieve this, it relies both in standardized technical components (connectors, applications, etc.) and in a strong certification program both for these components and for the entities that participate in the ecosystem (specially the operational environments). This way, it can be ensured that no "backdoors" are built into the software. Moreover, the security mechanisms are checked, as well, and they are made visible to other participants as a "trust level." Based on this information, everyone can decide whether the specified trust level of a potential partner is sufficient for their own use case or whether other partners need to be searched for.

To not create a singular and isolated perspective on the Data Spaces for manufacturing and Big Data, Boost 4.0 created a reference architecture (RA) [2]. The Boost 4.0 (RA) facilitates a common understanding within the Boost 4.0 project partners and supports the pilots in identifying Big Data assets required in order to fulfill their requirements. This model also brings together aspects from existing reference architectures and models targeting either Big Data (BDVRM [3], NBDRA) or the manufacturing domain (RAMI4.0, IIRA) and thus a clear alignment has been achieved. Also, the specific IDS standard was aligned with this RA and therefore the RA builds the foundation for the EIDS and gives the chance to put EIDS into perspective with many other reference architectures.

This chapter explains the basic design principles that were followed by Boost 4.0 to build the EIDS and presents some of the technical developments that enable the deployment of the EIDS. On the core of the EIDS there's the infrastructure that holds the entire data space, which is the main enabler for the data economy. On the next layer, there are the data space commonalities, which are the domain-specific specifications and structures. The outmost layer represents the raison d'être of the EIDS: this is where the data providers and the data users are represented, in the form of applications/platforms and specific use cases. Last, but not least, the EIDS

certification framework is the main source of trust, as it ensures the compliance of components and applications with the essence of the EIDS.

2   The European Industrial Data Space: Design Principles

The European Industrial Data Space builds its solid foundations on the IDS Infrastructure described in the IDS Reference Architecture Model. The value proposition of IDS can be summarized with three main pillars:

1. Unlimited **interoperability**, to enable connectivity between all kinds of different data endpoints.
2. **Trust**, to ensure a transparent and secure data exchange.
3. **Governance** for the data economy, which takes the form of usage control mechanisms and policy enforcement.

These three core pillars are put into practice in the so-called "Data Spaces" concept. This term refers to an approach to data management having the following characteristics:

- A federated data architecture. In fact, data do not need to be physically integrated. This is a crucial difference with central platforms, client service solutions, or cloud.
- No common schema. The integration will be at the semantic level and this does not affect domain-specific vocabularies.
- Data is an economic asset. In Data Spaces data is used as an economics asset whose value increases while you share the data.
- Ecosystem of Data Spaces. The power of Data Spaces lies in their scalability and nesting opportunities. There can be Data Spaces in Data Spaces, which are referred to as "ecosystem of Data Spaces," and this leads to large pools of cross-domain data.

The concept of Data Spaces combined with the pillars provided by IDSA makes sure to exploit the potential of the former, avoiding the lack of shared rules which is typical of the federated approach. Therefore, IDS is a crucial element to ensure data sovereignty, traceability, and trust among participants of a data space. The data space approach is foreseen in the European Strategy for Data of the European Commission, which suggests rolling out common European Data Spaces in nine strategic economic sectors. The manufacturing sector is one of them.

Boost 4.0 positions itself as a realization of such an embryonic data space built over IDS, to unlock the value of data treasures and Big Data services among the partners in its consortium, as displayed in Fig. 1. To achieve this, Boost 4.0 has deployed the required IDS infrastructure, has adopted the semantic model proposed by the IDS RAM which has been combined with different domain-specific standard vocabularies, and developed the certification framework that validates the implementation of the connectors that enable participation in the EIDS.
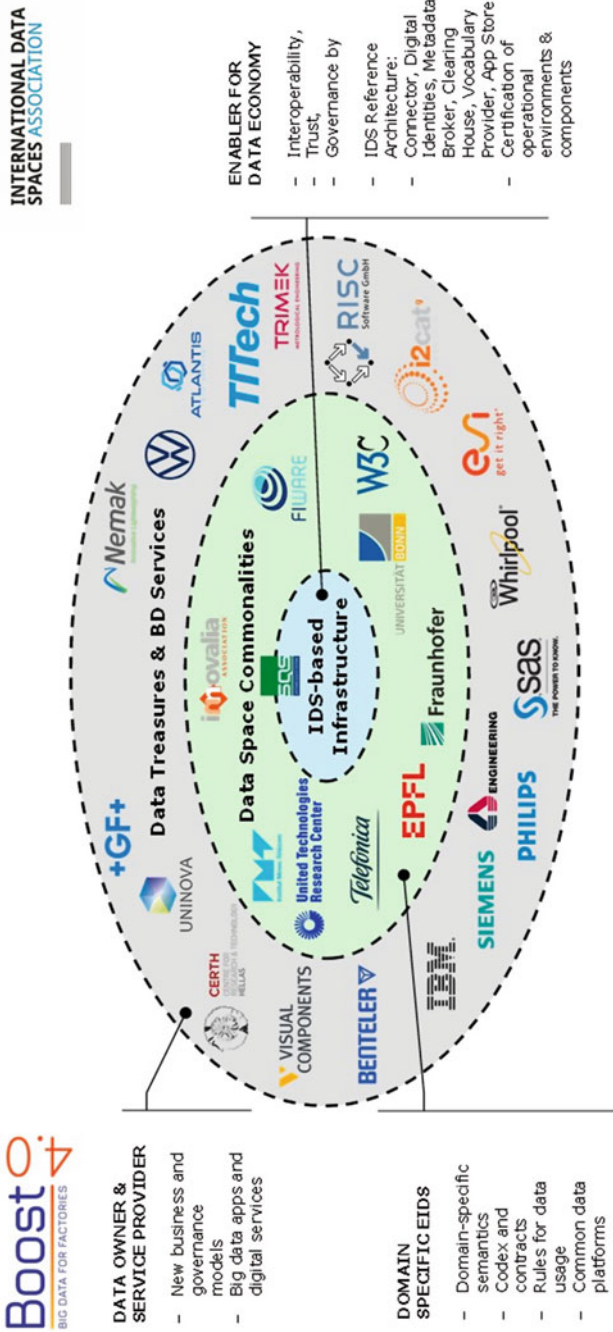
Fig. 1 Onion model for the European Industrial Data Space (EIDS)

Once the foundations of the EIDS are laid, it is necessary to adapt digital manufacturing platforms and to develop applications to provide an offering of services that attract data owners and other participants to the ecosystem.

3   EIDS-Based Infrastructure

The EIDS is based on the IDS Reference Architecture Model (RAM) V3.0 [4]. The IDS RAM defines how existing technologies must be orchestrated to create Data Spaces that deliver data sovereignty via a trusted environment to its participants. Among its five layers, the business layer defines a set of essential and optional roles for IDS-based Data Spaces. The essential components are the ones ensuring interoperability and a trustworthy data exchange thanks to clearly identifiable participants:

- The **connector** is defined as a dedicated communication server for sending and receiving data in compliance with the general connector specification (DIN SPEC 27070 [5]); different types of connectors can be distinguished (base connector vs. trusted connector, or internal connector vs. external connector).
- **DAPS** (Dynamic Attribute Provisioning Service), which issues dynamic attribute tokens (DATs) to verify dynamic attributes of participants or connectors.
- The **certificate authority** (CA), a trusted entity issuing digital certificates (X.509 certificates), may host services to validate the certificates issued.

The optional roles of IDS are the ones that make the data space more effective for discovering data endpoints, app search, and more:

- **Vocabulary Provider**. Data integration within IDS occurs at semantic level and this does not affect domain-specific vocabularies. IDS foresees vocabulary providers that help participants to discover domain-specific ontologies.
- **Metadata Broker**. This component is an intermediary managing a metadata repository that provides information about the data sources available in the IDS; multiple broker service providers may be around at the same time, maintaining references to different, domain-specific subsets of data endpoints.
- **App Store**. It is defined as a secure platform for distributing data apps; it features different search options (e.g., by functional or nonfunctional properties, pricing models, certification status, community ratings, etc.).
- **Clearing House**. This is an intermediary providing clearing and settlement services for all financial and data exchange transactions within the IDS.

Since the IDS RAM is technology agnostic for each data space that builds upon it, the challenge is to define its very own technology stack. The following paragraphs introduce some examples of technology stacks for some of the components that the EIDS uses.

3.1   HPC as Foundation for Resource-Intensive Big Data Applications in the EIDS

The Cirrus HPC service hosted at EPCC, the University of Edinburgh (UEDIN), provides both a high-performance computing (HPC) infrastructure and access to large storage systems. These infrastructures are suitable for Big Data analytic, machine learning, and traditional compute-intensive applications (e.g., CFD applications). The EIDS can access this infrastructure within the Boost 4.0 context.

The Cirrus [6] facility is based around an SGI ICE XA system. There are 280 standard compute nodes and 2 GPU compute nodes. Each standard compute node has 256 GiB of memory and contains two 2.1 GHz, 18-core Intel Xeon (Broadwell) processors. Each GPU compute node has 384 GiB of memory and contains two 2.4 GHx, 20-core Intel Xeon (Skylake) processors and four NVIDIA Tesla V100SXM2-16GB (Volta) GPU accelerators connected to the host processors and each other via PCIe. All nodes are connected using a single InfiniBand fabric and access the shared, 406 TiB Lustre file system. In addition to the Lustre file system that is included in Cirrus, it also has access to a high-capacity (i.e., in petabyte scale) object store system. This object store is similar to the Amazon S3 service and used to address medium- to longer-term data storage requirements.

In addition, for EIDS, a generic virtual machine (VM)-based infrastructure is deployed at EPCC to accommodate various infrastructure requirements where the above specialized infrastructure may not be suitable. Presently 10 VMs are available to Boost 4.0: each with 4 cores, 8 GB RAM, and 100 GB storage and runs Ubuntu version 18.04.

Furthermore, a certificate authority (CA) provided by the Hyperledger Fabric is deployed on the above generic EIDS VM infrastructure as described above. For more details about the Hyperledger Fabric CA, see the official documentation [7].

Additionally, a blockchain infrastructure based on the Hyperledger Fabric (v1.4) is deployed on the above EIDS VM infrastructure. Also, an Ethereum-based blockchain is deployed on this infrastructure. EIDS connectors could be linked, for example, to access these blockchain infrastructures.

3.2   Open-Source-Based FIWARE's EIDS Connector

During the runtime of Boost 4.0 FIWARE, in cooperation with the community members, Universidad Politécnica de Madrid (UPM) and Martel Innovate developed the EIDS CIM REST Connector. It can be implemented with the FIWARE Context Broker providing the NGSI interface in its proven version 2 or the latest ETSI GS CIM standard NGSI-LD embracing the concepts of Linked Data and the Semantic Web.

Both allow users to provide, consume, and subscribe to context information in multiple scenarios and involve multiple stakeholders. They enable close to

real-time access to and exchange of information coming from all kinds of data sources. Combined with FIWARE Smart Data Models, this approach offers easy and seamless interoperability within the IDS, while reusing the existing and well-adapted FIWARE standards and technology. Beyond that, also other existing REST(ful) APIs can be connected to the IDS.

The development of the FIWARE EIDS CIM REST Connector has been guided by several design goals. SME, developers, and interested parties shall be empowered to start quickly and easily into the world of FIWARE and IDS, with minimal upfront efforts but the full potential to scale up and adapt quickly. The technology stack shall be open, well-adapted, and cloud-ready. Service providers and diverse business models shall be supported with multi-cluster, multi-connector, multi-tenant, and multi-API options available from the very beginning to facilitate SME on-boarding, letting the EIDS grow. Four existing open-source projects were chosen to fulfill the ambitious design goals:

1. **Kubernetes** for automating the deployment, scaling, and management of containerized applications. As the current de facto industry standard, it groups containers that divide an application into logical units for easy management and discovery. Kubernetes can scale on demand and offers a vibrant community and an existing market for professional services.
2. **Rancher** for teams adopting Kubernetes management. It addresses the operational and security challenges of managing multiple Kubernetes clusters by offering extended security policies, user access management, Istio integration, and multi-provider and multi-cluster support. Fully featured tools for monitoring, logging, alerting, and tracing are provided right out of the box.
3. **Istio** as a service mesh to route and manage the traffic within the connector. It layers transparently onto existing distributed applications and microservices that can integrate into any logging platform or telemetry or policy system. Istio's diverse feature set lets you successfully, and efficiently, run a distributed microservice architecture and provide a uniform way to secure, connect, and monitor microservices, from simple to highly complex and restrictive scenarios.
4. **Ballerina** as an open-source programming language and platform for cloud-era application programmers. Ballerina includes the programming language and a full platform, which consists of various components of the surrounding ecosystem to offer a rich set of tools for integrated cloud development from coding to testing and deploying directly to the cloud.

As a result, the FIWARE EIDS CIM REST Connector offers a universal, unique, transparent, and fully integrated base-level approach to bring NGSI support to the IDS without the need for changes in application codes.

The generic approach of the FIWARE software components enables use cases ranging from Big Data, AI, or complex event processing to extended data analytics, data aggregation, and data anonymization, all combined with fine-grained access control down to the attribute level. Actionable insights can be conditionally propagated and notified to receiving devices and systems to leverage the highest value of near real-time data as best as possible. And last but not least, an innovative

approach for providing access and usage control to such publish/subscribe scenarios has been developed, incorporating the core trust concepts of the UCON model, the ORDL description language, and the XACML 3.0 specification.

4 Data Space Commonalities: Semantics as Prerequisite for Data Economy

Nowadays, vast amounts of data are produced daily by machines on assembly lines. These usually contain very useful information regarding manufacturing processes, and performing analysis might help offer precious insights to improve them. Unfortunately, when trying to achieve this goal, one is often confronted with the issue of data quality. Currently, in most companies, data exist in silos and data schemas underlying these systems are not harmonized. In concrete terms, this means that data comes in different formats, such as SQL databases, XML, Excel sheet, or CSV files. This implies that from one database or data file to another, the schemas can drastically change, since each service can decide on its own data schema and will select the one that better fits its needs. This means that from one service to another, the same piece of information can be attributed with different field names.

Another issue is the clarity of data schemas. To be comprehensible, data needs to follow a schema which helps associate the values to their meaning. Unfortunately, data schemas are not always easily understandable. Sometimes, the names of the fields in a data schema do not provide the user with helpful meaning. This is particularly prevalent in the case of legacy data, where sometimes fields had to follow a certain naming convention that did not coincide with natural language, making them particularly difficult to interpret in retrospect. Another issue that can happen with legacy data is when datasets had a fixed size, forcing users to reuse an existing field to store unrelated information, losing the meaning of the data in the process.

This creates a need within companies, but also within manufacturing domains at large, for a harmonized, universally understandable data schema. Ontologies can play a major role to help in that regard. The purpose of ontologies is to provide a shared common vocabulary among a domain context.

Thanks to ontologies, data can be marked up with meaningful metadata, which help to ensure that data is understood the same way across different parties. The different actors involved simply need to know the same ontology to understand the data. This leads to great time improvements, especially since concepts can easily be reused from one ontology to another. It also leads to increased interoperability, since concepts asserted in the ontology remain consistent in different applications. Whenever a piece of software uses the same ontology, the concepts and their relations cannot be misinterpreted.

This last point is another strong point in favor of ontologies: they help make data machine interpretable. Thanks to metadata tags, a piece of code can easily read the data, understand what it stands for, and treat it accordingly.

4.1   The IDS Information Model

Exchanging data and operating the infrastructure of an ecosystem that supports such exchange, both require a common language. The IDS Information Model [8, 9] is the common language used for:

- Descriptions of digital resources offered for exchange.
- The self-description of infrastructural components, such as connectors, providing data or metadata brokers enabling potential consumers to find them.
- The headers of the messages that such components send to each other.

The IDS Information Model is conceptually specified in Sect. 3.4 "Information Layer" of the IDS RAM as human-readable text and UML diagrams. Its declarative representation as an ontology [10] based on the World Wide Web Consortium's family of Semantic Web standards around the RDF Resource Description Framework provides an unambiguous machine-comprehensible implementation for the purpose of validation, querying, and reasoning. Finally, a programmatic representation addresses the purposes of integration into services, tool support, and automated processing [11].

The IDS Information Model addresses all concerns of sharing digital resources, i.e.,

- Their content and its format and structure.
- Concepts addressed by a resource.
- The community of trust, whose participants exchange digital resources.
- Digital resources as a commodity.
- The communication of digital resources in a data ecosystem.
- The context of data in such a resource.

As shown in Fig. 2, the conceptualization and implementation of most of the concerns addressed by the IDS Information Model builds on established standards, key pillars being an extension of the W3C Data Catalog Vocabulary (DCAT) [12] and a usage policy language based on the W3C Open Digital Rights Language (ODRL) [13].

Depending on the requirements of each concern, the Information Model's declarative representation as an ontology literally reuses existing standards, or it adapts and extends them to the specific requirements of data ecosystems.
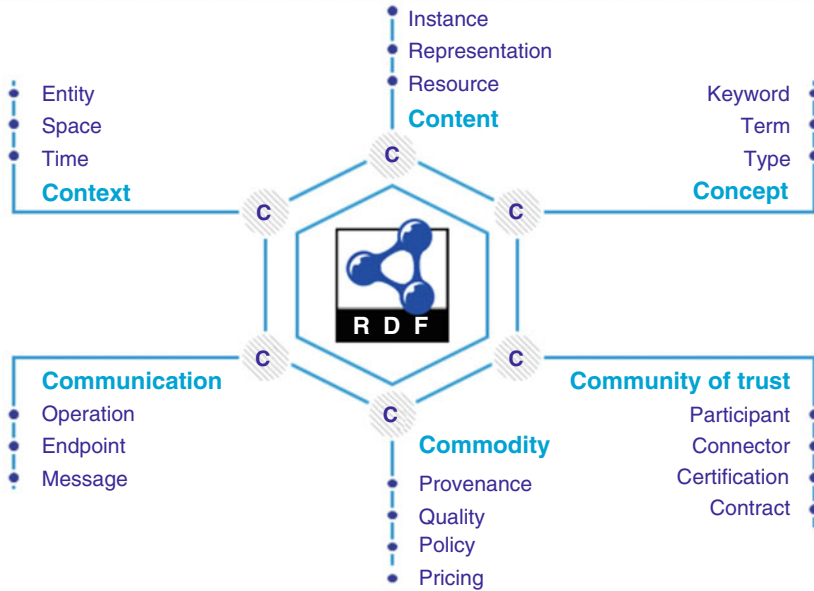
Fig. 2 The Concern Hexagon of the IDS Information Model in detail, with references to standards built on IDS Reference Architecture Model

4.2 Domain-Specific Ontologies in EIDS: QIF

The IDS Information Model takes a domain-agnostic perspective on data. To reasonably exchange data in specific domains such as manufacturing in the context of the EIDS, a connection with domain ontologies is required. While the IDS mandates semantic descriptions of data resources by metadata in terms of the IDS Information Model for interoperability, it does not enforce but strongly encourages the use of standardized, interoperable representations for the actual data as well. In other words, data should use a standardized domain ontology as their schema. The IDS Information Model bridges the domain-agnostic metadata level and the domain-specific domain knowledge with regard to multiple aspects of the content and concept concerns. It is recommended to use structured classification schemes, e.g., taxonomies, instead of just string-based keywords to characterize the concepts covered by a digital resource. The content of a digital resource in manufacturing may, for example, be sensor measurements expressed using the W3C Sensor, Observation, Sample, and Actuator/Semantic Sensor Network (SOSA/SSN) ontology [14]. The bridge to the metadata level is established by using, for example, the W3C Vocabulary of Interlinked Datasets (VoID) [15] to express that the digital resource mainly contains instances of the class *Sensor*, or using the W3C Data Cube Vocabulary [16] to express that the digital resource consists of a three-

dimensional matrix with temperature measurements in the dimensions: (1) time, (2) geo-coordinates, and (3) sensor used.

One example, used in Boost 4.0, of a domain-specific standard ontology is the Quality Information Framework (QIF). QIF is a feature-based ontology of manufacturing metadata, built on XML technology, with the foundational requirement of maintaining traceability of all metadata to the "single source of truth"—the product and all its components as defined in CAD/MBD. It is an ISO 23952:2020/ANSI standard which includes support for a complete semantic derivative model-based definition (MBD) model, measurement planning information, and measurement result information. This characterization allows users to establish a Digital Twin by capturing the duality of aspects of manufacturing data: the as-designed product and the as-manufactured product – including the mappings between the two aspects.

QIF is a highly organized grammar and vocabulary for software systems to communicate manufacturing data structures. With software interoperability as the goal, vendors and end-users were available to verify (through both pilot studies and production-system deployments) the robustness of QIF and its truly semantic structure. Another important design objective is data traceability to the authority product definition model. Each piece of metadata in a QIF dataset is mapped to the owning GD&T (geometric dimensioning and tolerancing), feature, and model surface in the authority CAD model. This high-resolution mapping to the "single source of truth" ensures that any other data derived from the model can be traced through a graph to the mass of QIF data. This is why QIF matters beyond just metrology: it provides the mapping to metrology data for the entire model-based enterprise.

The software vocabulary and grammar of QIF is defined by the QIF XML Schema Definition Language (XSDL) schemas. These schemas are at the core of what defines QIF. At a high level, there are three primary pillars to QIF's approach to these data quality: XSD validation, XSLT integrity checks, and digital signatures. XML Schema Definition (XSD) validation is a test where a QIF instance file is checked for validity to the QIF digital "language" defined in the QIF schemas. Extensible Stylesheet Language Transformations (XSLT) is a Turing-complete language for processing XML files. Included with the QIF standard is a set of XSLT scripts which can carry out integrity checks on QIF instance files that are not possible with a simple XSD validation test. QIF has the ability to control the provenance of the data via digital signatures. This infrastructure helps to ensure that QIF instance files can be imbued with the trust necessary to use it throughout the entire product life cycle.

5　Data Treasures and Big Data Services in EIDS: Big Data Applications and Use Cases

This section describes in high level the types of services, applications, and platforms that are available in EIDS ecosystem and some examples that have been deployed in the Boost 4.0 project, each of which focuses on the integration with one of the design principles, although obviously they mostly integrate several of them.

The Boost 4.0 project provides plenty of services and tools for the application of Big Data technologies at different stages of the factory life cycle and the supply chain (cf. Table 1). The project contributes on the state of the art in cognitive

Table 1 Big Data solutions that are made available in the European Industrial Data Space (EIDS)

Platform/app	Key features
RISC Data Analysis Platform for the Optimization of Flexible Machines Manufacturing	Combination of BD technologies and semantics to process and store data. It is based on cluster computing and data analytics services
ESI Hybrid Twin for Zero Defect Production	Cloud platform dedicated to simulation solutions. Provides 3D remote visualizations and a vertical application framework
Interactive Analytics for Smart injection Molding	Live monitoring of the machine's behavior. Graphical interface for mapping data with concepts of a generic data model. Outlier detection
ESI Platform for Cognitive Value Chain	Big Data and visual analytics. Use of ontology for the extraction of reliable training datasets for predictive maintenance
Visual Components 4.0 and UNINOVA Analytics for Inventory Optimization and Dynamic Production Planning	BD infrastructure for data management, storage, and processing. Visualization services based on Grafana. Data aggregation into a 3D Digital Twin by exploiting the connection with Visual Components Simulation Software
Siemens MindSphere Extensions for third Parties	Cloud-based IoT platform for BD management and analysis with application in various domains
IBM Blockchain Application for Certification of Origin	A blockchain-based application for enabling traceability and full visibility of transaction data related to the production of fabrics
CERTH Cognitive Analytics Platform for Industry 4.0	Cloud platform for BD analytics and advanced visualization. Real-time model retraining and ensemble methods for predictive maintenance. EIDS compatible
SAS Analytics Platform for Production Planning	Powerful analytics platform with plenty of real-world applications. Support of extensions through web services
ATLANTIS DSS and Fraunhofer IEM Analytics for Predictive Maintenance	Complete solution that provides data analytics services alongside with decision support mechanisms. EIDS compatible
TRIMEK M3 Platform for Zero Defect Manufacturing	Metrology platform that provides BD analysis and visualizations for scanning processes
I2CAT IoT Real-Time Location System	IoT location monitoring system with high precision (1–2 cm). EIDS compatible

manufacturing using the latest available techniques and methodologies in Big Data management, processing, modeling, analysis, and visualization. More precisely, Big Data services, tools, and platforms related to the detection of deterioration rate of production machines and their root cause are available [17]. Furthermore, predictive and cognitive modeling methodologies and early anomaly detection algorithms are available as well through the EIDS infrastructure [18]. The list of provided analytics services is completed by the implemented planning optimization algorithms and advanced data visualization and visual analytics tools [19]. In addition to these solutions, services and apps related to the connection of Big Data analytics platforms with data sources from shopfloor and other tools are also provided. These services/apps are EIDS connector-compliant application for specific analytics platforms that participated in the project and they are major players in analytics services in global market at the same time. Table 1 presents an overview of the solutions for which the compatibility was determined during the project runtime. A selection of these solutions is described within this section.

5.1 Types of Connection for Big Data Analytics Services and Platforms in the EIDS

There are two main ways that Boost 4.0 analytics services or platforms can be made available through the EIDS enabled ecosystem as it is depicted in Fig. 3.

1. An analytics service or platform can be connected to the EIDS by using its own connector. This type of connection fits better for Big Data platform integration with EIDS as these platforms have a lot of dependencies and IPR restrictions so it is difficult to be packaged and deployed as apps in order to be available for downloading and execution in an IDS connector.
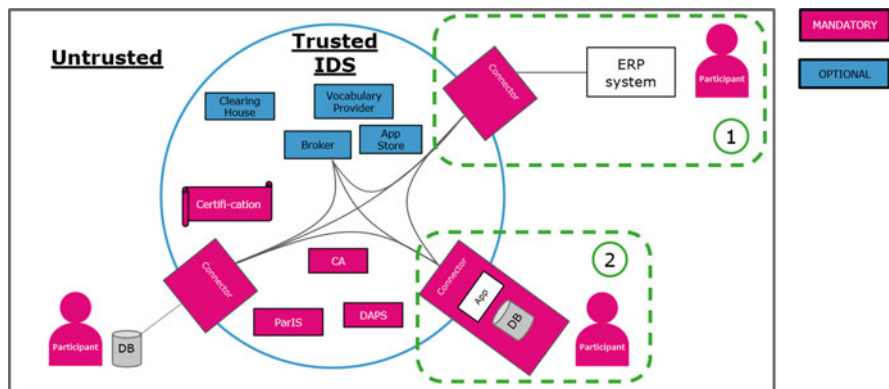

Fig. 3 Connection of EIDS ecosystem with Boost 4.0 Big Data analytics apps/platforms

2. Analytics method/service/app can be deployed by using microservice concepts (and Docker containers) and be compiled internally into a connector. This solution fits better for stand-alone applications and algorithms as it is easier to be compiled and executed in a connector.

5.2 Example of Integration of an Open-Source-Based Connector Within a Big Data Analytics Platform Connected: The CERTH Cognitive Analytics Platform

5.2.1 Manufacturing Scenario

The fast growth of data creation and gathering from a wide range of sources in factories and supply chains led to significant challenges in data analytics solutions, data collection, and handling. Big Data availability boosts predictive maintenance solution for smart factories. However, the data availability itself cannot create a self-learning factory that is able to continuously learn and act. This Big Data platform example presents a solution that aims to solve both problems of self-learning and trusted data connection and transfer.

5.2.2 Digital Business Process

The Cognitive Analytics platform for Industry 4.0 is one of the EIDS Big Data applications of Boost 4.0. The Cognitive Analytics platform exploits the Big Data of factories and feeds them to machine learning algorithms aiming to enhance industrial automation through real-time fault diagnosis and production planning optimization. It is an end-to-end user platform supporting predictive maintenance functionalities, and it is delivered as a Big Data application compatible with the EIDS ecosystem of Boost 4.0 as it is able to provide and consume data by using the IDS trusted connector. Figure 4 depicts the functionalities and achievements of the Cognitive Analytics platform.

Two core functionalities are provided within the platform: anomaly detection for predictive maintenance and production scheduling optimization. The anomaly detection feature of the platform receives multivariate time series data coming from machines. Historical data are stored in the platform's database and live streaming data are coming through the IDS connector. The platform provides online training of historical data, live monitoring of real-time data, and automatic cognitive retraining which keeps the system's performance at high levels. The EIDS application of production scheduling optimization is demonstrated on a real-case scenario of a fabric production's weaving process. The main goals of the scheduling optimization of the weaving process are finishing orders by their deadline, cost-efficient production, and prioritization of orders. Based on advanced visualization
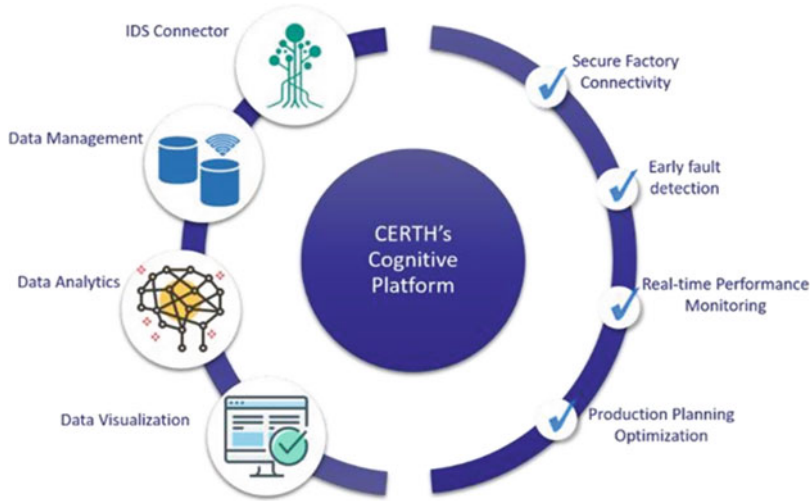
Fig. 4 The functionalities of the Cognitive Analytics platform in conjunction with its high-level achievements

techniques, a monitoring user interface was developed for this application showing the status and calculating the arrangement of orders to machines in real time.

5.2.3 EIDS Integration Approach

The open-source-based IDS connector accomplishes the integration of real-time machine data with the Cognitive Analytics. Secure Big Data exchange is achieved with the configuration of IDS trusted connector ecosystem. The system encompasses two IDS trusted connectors. The first connector is deployed on the factory site. The IDS connector receives the machines' data through an MQTT broker which is connected with the factory's cloud infrastructure. The factory cloud repository is the data provider of IDS architecture. The second IDS connector is placed in the data consumer site, specifically the Cognitive Analytics framework. A MQTT broker is also used to enable data exchange with the data provider. The architecture accomplishes secure and trusted communication between all system components, especially in the case of sensitive and private data of factories.

5.3 Example of Integration of the Open-Source-Based Connector with Manufacturing Supermarkets 4.0: Volkswagen Autoeuropa

5.3.1 Manufacturing Scenarios

Data collection, harmonization, and interoperability are difficult to achieve in all datasets, but in Big Datasets it is even worse, due to the sheer volume of the data to be transformed. Volkswagen Autoeuropa is an automotive manufacturing industrial plant located in Portugal (Palmela) since 1995 and part of Volkswagen Group. Currently, the logistics process is heavily reliable on manual processes and each step of the process creates data that is stored in a silo-based approach. In order to develop a true Digital Twin of the logistics process, there is the need to integrate all data silos and to make them interoperable between themselves and with other, external data sources, so as to enable real-world data provision to the Digital Twin.

5.3.2 Digital Business Process

The solution for this issue is based on a dedicated IDS-supported Big Data application that enables data harmonization and integration across the different data silos, by collecting and harmonizing the data from each silo into a common database system. The proposed Big Data app addresses the following technical requirements: (1) able to deal with raw data in many formats and sizes; (2) assure data quality; (3) efficient Big Data transformation and storage; (4) being able to address interoperability at the data level, enabling the development of additional added value services for users; (5) inclusion of custom schemas, in order to transform and harmonize data into standardized or proprietary schemas; and (6) a robust and efficient distributed storage system that is scalable in order to process data from data sources.

5.3.3 EIDS Integration Approach

The main open-source tools used for developing the proposed architecture were (1) Apache Spark, used for large-scale data processing, which includes the tasks of data cleaning and transformation; (2) Redis, as a NoSQL in-memory approach, for storing and managing raw data; and (3) PostgreSQL as the final database system that stores harmonized data (PostgreSQL could be replaced by any other database system). An IDS trusted connector is used to access raw, unharmonized data within the IDS ecosystem, and another IDS connector is used to publish harmonized data back to the IDS ecosystem. The fact that the IDS ecosystem is compliant with the containerization approach adopted for the Big Data app, in this case using

Docker Swarm orchestration, is a big advantage in terms of integration with the IDS ecosystem and with the IDS App Store.

5.4 Example of Integration of the Domain-Specific Ontologies with Predictive Maintenance Processes: OTIS

5.4.1 Manufacturing Scenarios

OTIS manufacturing system for elevator panel production line generates multiple data silos coming from MES, machines, and ERP systems. Dispersed information is difficult to integrate, thus rendering process to run in local optima. Aggregated data will have significant impact on overall manufacturing process improvements using data integration, analytics, and modeling. Inside the process team looked at optimization of the following adverse effects observed in the process:

- Maintenance cost reduction—due to various mechanical breakdowns or incorrect maintenance during production process. This results in production stop and higher machine maintenance cost to resume production processes.
- OEE (Overall Equipment Effectiveness)—increase equipment operation time vs. maintenance time.
- Discover hidden causes of production stops—combining distributed data silos and performing dataset mining.

5.4.2 Digital Business Process

Envisioned and realized data aggregation and integration using Boost 4.0 technologies enables production process to run more optimally. The solution consists of two parts: (1) Bayesian causal model that describes details of manufacturing processes derived from data mining and analysis and (2) information aggregation and sharing with supply chain via FIWARE IDS connector to enable global-level production optimization.

5.4.3 EIDS Integration Approach

The replication pilot final version is using two technologies to aggregate data coming from production sensors. On local level MQTT broker is used to aggregate data from machine sensors and systems. On Boost 4.0 level pilot used FIWARE IDS connector that integrates with MQTT broker via plugin to share production information with Boost 4.0 consortium. The usage of domain-specific vocabularies is extremely important in this example, as it deals with diverse sources, with different machine providers and sensors of all kinds.

5.5 Example of Integration of EIDS-Based Infrastructure with Warehouse Management Processes—Gestamp

5.5.1 Manufacturing Scenarios

Gestamp's participation in Industry 4.0 initiatives aims to create more efficient, consistent, and reliable manufacturing plants by adding intelligence to the production processes and getting the right information to the right people. These efforts are often hindered by the unavailability of precise, fine-grained, raw data about these production processes. Moreover, the skills required to apply advanced predictive analytics on any available data might not be present inside the company, which entails the need to securely share the data with expert third parties.

5.5.2 Digital Business Process

These issues have been addressed by providing a state-of-the-art indoor real-time locating system (RTLS). The system deploys IoT sensors on the plant shop floors to gather raw data directly from key mobile assets involved in the manufacturing processes, such as coils, dies, semi-finished and finished products, containers, forklifts, and cranes. Seamlessly integrating these IoT sensors with the IDS ecosystem, the solution grants access to the gathered data from outside of Gestamp's in-house network only through secure IDS connectors, thus facilitating data sharing while simultaneously adhering to Gestamp's strict security restrictions.

This previously untapped data can then be exploited to increase the logistical efficiency of the production processes by:

- Improving operational effectiveness.
- Increasing flexibility of production.
- Allowing more dynamic allocation of resources.
- Reducing changeover time.
- Refining warehouse and storage area management.

5.5.3 EIDS Integration Approach

The RTLS consists of ultra-wideband (UWB) tags and anchors, radio frequency identification (RFID) tags, and the proprietary *i2Tracking* stack. Other technologies used in the system include MQTT message brokers, Python-based data processors, and MongoDB and SQL databases. The design and technologies of the system and the integration in the EIDS infrastructure guarantee that the collected real-time data is made available in the IDS space with minimal overhead, which allows for more robust and precise analysis of the data. The solution also includes sample IDS consumer applications that show how the provided data can be ingested and utilized by third-party consumers.

5.6   Example of Integration of EIDS-Based Infrastructure with Logistics Processes: ASTI

5.6.1   Manufacturing Scenarios

Due to the special features of 5G networks such as high availability, ultra-low latency, and high bandwidth, Industry 4.0 proposes the use of this technology to support the intra-factory communications in replacement of the current communication practices mainly based on WLAN (IEEE 802.11 family). 5G networks, in addition to improved transmission capabilities, include the allocation of the computational resources closer to the factories for reducing latencies and response times.

Furthermore, the use of Artificial Intelligence and machine and deep learning techniques is substantially boosting the possibilities for prediction of complex events that help to take smart decisions to improve the industrial and logistic processes.

5.6.2   Digital Business Process

In this context, an interesting use case that combines Industry 4.0, 5G networks, an IDS trusted connector, and Artificial Intelligence and deep learning techniques is proposed. By this combination it is possible to predict the malfunctioning of an automated guided vehicle (AGV) connected through 5G access with its PLC controller deployed and virtualized in a multi-access edge computing (MEC) infrastructure, by exclusively using network traffic information and without needing to deploy any meter in the end-user equipment (AGV and PLC controller).

5.6.3   EIDS Integration Approach

Intensive experiments with a 5G real network and an industrial AGV in the 5TONIC [20] environment validate and prove the effectiveness of this solution. By using deep neural networks, and only analyzing the network parameters of the communication between the AGV and the PLC controller, several time series are built based on 1-D convolutional neural network (CNN) models that are able to predict in real time that the AGV is going to lose its trajectory 15 s ahead, which allows taking pre-emptive actuations. An IDS trusted connector acts as a bridge to transmit the CNN prediction outside the MEC infrastructure to an external dashboard based on the Elasticsearch-Logstash-Kibana (ELK) stack.

6   Certification as Base for Trust in the EIDS

Data security and data sovereignty are the fundamental value propositions of the EIDS. Any organization or individual seeking permission to access the EIDS must

certify the core components, like connectors, to securely exchange data with any other party which is part of the data space.

The EIDS components are based on the International Data Space Reference Architecture Model V3.0, which also defines a certification criteria catalogue. Both Data Spaces, IDS and EIDS, are referring to the same criteria catalogues. The catalogue is split into three thematic sections, i.e., IDS-specific requirements, functional requirements that are taken from ISA/IEC 62443-4-2, and best practice requirements for secure software development.

The EIDS core components must provide the required functionality and an appropriate level of security. Therefore, the IDS certification scheme defines three security profiles for the core components:

- **Base Security Profile**: includes basic security requirements: limited isolation of software components, secure communication including encryption and integrity protection, mutual authentication between components, as well as basic access control and logging. However, neither the protection of security-related data (key material, certificates) nor trust verification is required. Persistent data is not encrypted and integrity protection for containers is not provided. This security profile is therefore meant for communication inside of a single security domain.
- **Trust Security Profile**: includes strict isolation of software components (apps/services), secure storage of cryptographic keys in an isolated environment, secure communication including encryption, authentication and integrity protection, access and resource control, usage control, and trusted update mechanisms. All data stored on persistent media or transmitted via networks must be encrypted.
- **Trust + Security Profile**: requires hardware-based trust anchors (in the form of a TPM or a hardware-backed isolation environment) and supports remote integrity verification (i.e., remote attestation). All key material is stored in dedicated hardware-isolated areas.

Within the Boost 4.0 project, the Spanish company SQS has developed the following infrastructures to test IDS and EIDS components.

6.1   IDS Evaluation Facility

SQS has defined a test laboratory (Lab. Q-IDSA) that integrates infrastructures already available in SQS quality as a service (QaaS) offer with new developments and processes required to validate IDS components. It will be accredited by IDSA and has also the scope of being ISO17025 accredited, which will make the lab a test center.

To carry on the certification process, SQS has defined a set of activities that imply technical assessment, where functional, interoperability, and security testing is performed, and documentation and processes are reviewed. As such, the adequacy and completeness of the installation and operational management documents are

judged, and the adequacy of the security procedures that the developer uses during the development and maintenance of the system is determined.

A detailed description of the evaluation process of IDS-based components can be found in the position paper "IDS Certification Explained" [21].

6.2 Integration Camp

SQS has built an architecture with real IDSA components with the goal of having a full IDSA environment. The architecture was first built with minimum components needed to test the interoperability of connectors and base of IDSA environment, and it is in constant evolution, including more components (i.e., DAPS, Broker), building an architecture where every IDS component can be tested.

This architecture is opened for everyone, in a monthly event (Fig. 5), as a remotely accessible infrastructure where participants can test if their components are ready to work in a real IDS environment. The evaluation facility is the ideal place for those who want to prepare their IDS connectors and other IDS components for certification.
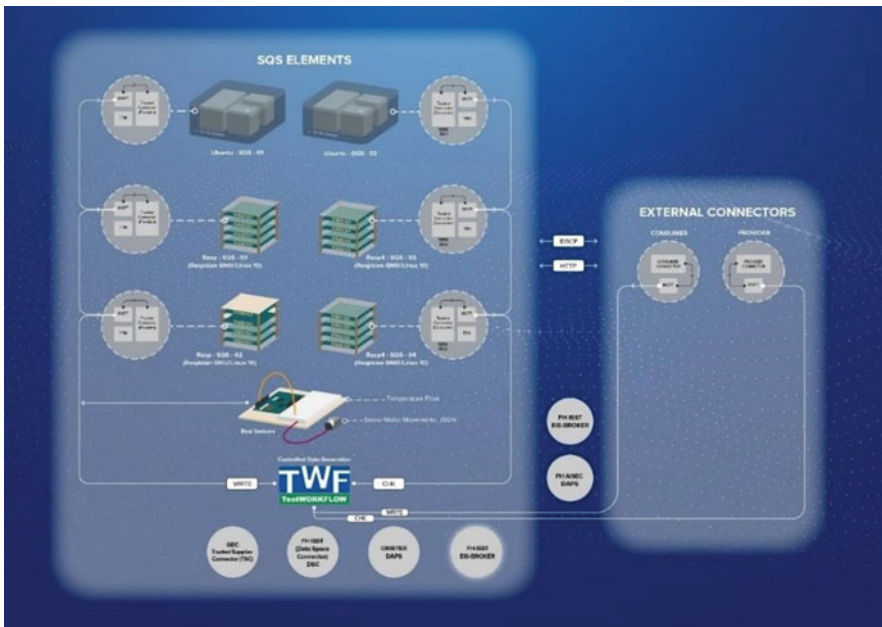
Fig. 5 Architecture of the integration test camp—as of the fourth iteration (Nov)

7 Data Space for the Future of Data Economy

This chapter gave a high-level overview on how Data Spaces are designed in the context of the IDS Reference Architecture and standards and on how the Boost 4.0 project used it in order to come up with its very own embryonic data space for the manufacturing domain—the European Industrial Data Space.

The fundamental design principle for Data Spaces must fulfill three core functionalities, all of which were encountered in the IDS-based EIDS ecosystem:

1. **Interoperability.**
2. **Trust.**
3. **Governance.**

Figure 6 puts these three pillars into context with generic reference models, domain-specific aspects, and business case-specific enablers. This figure shows all topics which Boost 4.0 dealt with, like domain-specific standards and formats, metadata, exchange of data, identification and authentication, authorization, certification and monitoring, and governance. Furthermore, it shows the topics that are still to be solved yet and must be encountered in future projects. Among those topics, like authorization and usage control, legal and operational agreements will play a significant role.

The creation of a data space for manufacturing is a milestone which is of tremendous importance for the future of data economy, since it brings data sovereignty to those who own the data treasure and therefore helps to break data silos and leverage its value. The fact that the manufacturing domain is one of four core domains that have been declared as the focus areas by the European Commission in the Open DEI [22] project (besides health, energy, and agri-food) shows how meaningful the development and exploration of the EIDS is. Besides, there are further first projects and initiatives that are aiming at the goal to come up with Data Spaces, either for a specific domain or with a broader scope:
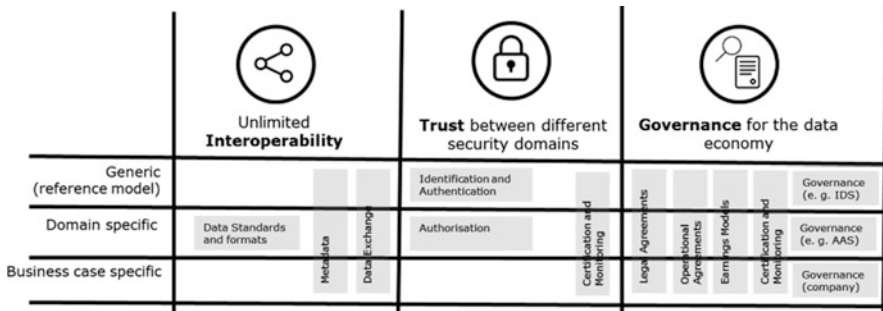

Fig. 6 Three pillars for Data Spaces and their enablers

- Gaia-X [23] that is a federated data infrastructure for Europe based on IDS in the field of data sovereignty.
- The IDS Launch Coalition [24] that focuses on creating an offering of.
  IDS-based products for the market
- The German "Datenraum Mobilität" [25] (eng. Mobility Data Space), also based on IDS and focusing on the mobility domain.

A growing amount of use cases that do not only connect two data endpoints but rather show the potential of Data Spaces, like the EIDS, by supporting many-to-many connections is on its way to change our perception of how data will be treated and traded in the future.

References

1. Boost 4.0. https://boost40.eu
2. Boost 4.0. *Public Deliverables*. http://boost40.eu/deliverables/
3. Curry, E., Metzger, A., Berre, A. J., Monzón, A., & Boggio-Marzet, A. (2021). A reference model for big data technologies. In E. Curry, A. Metzger, S. Zillner, J.-C. Pazzaglia, & A. García Robles (Eds.), *The elements of big data value* (pp. 127–151). Springer International Publishing.
4. International Data Spaces Association. (2019). *Reference architecture model*. Available at: https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-ReferenceArchitecture-Model-3.0.pdf
5. International Data Spaces Association. (2020). *IDS is officially a standard: DIN SPEC 27070 is published*, February 21 2020. https://internationaldataspaces.org/ids-is-officially-a-standard-din-spec-27070-is-published/
6. *Cirrus Documentation*. https://cirrus.readthedocs.io/en/master/index.html
7. Hyperledger Fabric CA Documentation. https://hyperledger-fabric-ca.readthedocs.io/en/release-1.4/
8. IDSA Tech Talk on 3 September 2020; cf. https://www.internationaldataspaces.org/recordedidsa-live-sessions/#techtalks
9. Bader et al. (2020). The international data spaces information model – an ontology for sovereign exchange of digital content. *International Semantic Web Conference*. Springer.
10. Available at https://w3id.org/idsa/core, source code maintained under the Apache 2.0 license at https://github.com/International-Data-Spaces-Association/InformationModel/
11. *Java implementation*. Available at https://maven.iais.fraunhofer.de/artifactory/eis-ids-public/de/fraunhofer/iais/eis/ids/infomodel/
12. *W3C Data Catalog Vocabulary DCAT*: https://www.w3.org/TR/vocab-dcat-2/
13. *W3C Open Digital Rights Language ODRL*: https://www.w3.org/TR/odrl-model/
14. *W3C Sensor, Observation, Sample and Actuator/Semantic Sensor Network (SOSA/SSN) ontology*: https://www.w3.org/TR/vocab-ssn/
15. *W3C Vocabulary of Interlinked Datasets VoID*: https://www.w3.org/TR/void/
16. *W3C Data Cube Vocabulary*: https://www.w3.org/TR/vocab-data-cube/

17. Lázaro, O., Alonso, J., Figueiras, P., Costa, R., Graça, D., Garcia, G., et al. (2022). Big data driven industry 4.0 service engineering large scale trials: The Boost 4.0 experience in technologies and applications for big data value. In E. Curry, S. Auer, A. J. Berre, A. Metzger, M. S. Perez, & S. Zillner (Eds.), *Technologies and applications for big data value* (pp. 373–397). Springer.

18. Lázaro, O., Alonso, J., Holom, R. M., Rafetseder, K., Kritzinger, S., Ubis, F., et al. (2022). Model based engineering and semantic interoperability for trusted digital twins big data connection across the product lifecycle in technologies and applications for big data value. In E. Curry, S. Auer, A. J. Berre, A. Metzger, M. S. Perez, & S. Zillner (Eds.), *Technologies and applications for big data value* (pp. 399–429). Springer.

19. Lázaro, O., Alonso, J., Ohlsson, P., Tijsma, B., Lekse, D., Volckaert, B., et al. (2022). Next-generation big data-driven factory 4.0 operations and optimization: The Boost 4.0 experience in technologies and applications for big data value. In E. Curry, S. Auer, A. J. Berre, A. Metzger, M. S. Perez, & S. Zillner (Eds.), *Technologies and applications for big data value* (pp. 345–371). Springer.

20. *5Tonic, an Open Research and Innovation Laboratory Focusing on 5G Technologies*: https://www.5tonic.org/

21. *IDSA position paper Certification Explained v 1.0*: https://www.internationaldataspaces.org/wp-content/uploads/2020/09/IDSA-Position-Paper-IDS-CertificationExplained.pdf

22. *OPEN DEI*: https://www.opendei.eu/

23. *Gaia-X*: https://www.data-infrastructure.eu/

24. *IDSA Launching Coalition*: https://www.internationaldataspaces.org/the-launching-coalition-start-of-a-new-journey-on-sovereign-data-sharing/

25. Delhaes, D. (2020-10-28). *Merkel drängt Autokonzerne: BMW, Daimler und VW sollen Datenschatz teilen [Merkel urges car companies: BMW, Daimler and VW should share a wealth of data]*. Available at: https://www.handelsblatt.com/politik/deutschland/autogipfelmerkeldraengt-autokonzerne-bmw-daimler-und-vw-sollen-datenschatz-teilen/26308418.html