Report from Dagstuhl Seminar 19391

Data Ecosystems: Sovereign Data Exchange among Organizations

Edited by

Cinzia Cappiello[1], Avigdor Gal[2], Matthias Jarke[3], and Jakob Rehof[4]

1 **Polytechnic University of Milan, IT,** cinzia.cappiello@polimi.it
2 **Technion – Israel Institute of Technology – Haifa, IL,** avigal@technion.ac.il
3 **RWTH Aachen, DE,** jarke@dbis.rwth-aachen.de
4 **TU Dortmund, DE,** jakob.rehof@cs.tu-dortmund.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 19391 "Data Ecosystems: Sovereign Data Exchange among Organizations". The goal of the seminar was to bring together people from different disciplines (also outside the computer science area), in order to identify (i) a set of research challenges for the future development of data ecosystems and a catalogue of major approaches relevant to the field and (ii) a set of developed use cases of particular interest to the further development of data ecosystems. Towards the objectives, the seminar included tutorials, invited talks, presentations of open problems, working groups. This report presents the most relevant findings and contributions.

Seminar September 22–27, 2019 – http://www.dagstuhl.de/19391
2012 ACM Subject Classification Information systems → Information integration, Information systems → Data exchange, Information systems → Ontologies, Computing methodologies → Machine learning, Computer systems organization → Architectures, Security and Privacy, Information systems → Data analytics, Software and its engineering, Social and professional topics → Computing and business
Keywords and phrases Data sovereignty, Data ecosystems, Business models, Data integration, Ethics
Digital Object Identifier 10.4230/DagRep.9.9.66
Edited in cooperation with Bernadette Farias Lóscio (Collector)

1 Executive Summary

Cinzia Cappiello (Politecnico di Milano, IT)
Avigdor Gal (Technion – Haifa, IL)
Matthias Jarke (RWTH Aachen, DE)
Jakob Rehof (TU Dortmund, DE)

The design of *data ecosystems*, infrastructures for the secure and reliable data exchange among organizations, is considered as one of the key technological enablers for digitalization and the digital economy of the future. Several applied research initiatives and industry consortia provide substantive evidence of this trend e.g., the Industrial Internet Consortium (IIC)[1]

[1] https://www.iiconsortium.org/

Data Ecosystems: Sovereign Data Exchange among Organizations, *Dagstuhl Reports*, Vol. 9, Issue 9, pp. 66–134
Editors: Cinzia Cappiello, Avigdor Gal, Matthias Jarke, and Jakob Rehof
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

formed in the USA, the Industrial Data Space (IDS) founded in Germany and the associated consortium International Data Space Association (IDSA)[2]. Most of these initiatives aim to provide a *reference architecture* for dealing with (i) *governance* aspects related to the definition of policies and conditions able to norm the participation to the data ecosystem, (ii) *security* aspects related to the definition of policies and infrastructures for guaranteeing a trusted and secure exchange of data, (iii) *data and service management* aspects related to representation models and exchange formats and protocols, and (iv) *software design* principles related to the realization of the architectural components and their interaction.

All these aspects have been discussed in the seminar and the main findings are described in this report. In addition, a central new aspect of data ecosystems that we considered in the seminar lies in the view of data as having an economic value next to its intrinsic value to support operational and decisional core business activities. This means that in the data ecosystem, data is typically considered both a business asset and a business commodity which may be priced and sold in some form (e.g., data provisioning service or raw data) according to contracts.

As testified by the amount and variety of problems described above, the creation of such ecosystems poses many challenges cutting across a wide range of technological and scientific specializations. For this reason, the seminar involved researchers from different communities. Interdisciplinary discussions gave the possibility to analyze different perspectives and to achieve valuable outcomes presented in this report, such as a wide set of research challenges and the definition of interesting use cases for the further development of data ecosystems. Details about the activities carried out during the seminar are provided in the following.

Overview of the activities

The seminar took place from Monday September 23 until Friday September 27. The seminar program encompassed four invited talks (keynotes and tutorials) on the first day (Sep. 23rd), by Gerald Spindler (law and ethics), Frank Piller (ecosystems and business models), Maurizio Lenzerini (data integration), and Boris Otto (International Data Space). After discussions related to the talks and tutorials, the remaining afternoon was spent structuring (through joint discussion) the coming days of the seminar and group structure. As a result, group structure was based on a thematic structure encompassing three groups, one for each of the topic areas Business, Data, and Systems. Tuesday Sept. 24 began with a breakout into groups and election of scribes in each of the three groups (Business, Data, and Systems), and the remainder of the day was taken up by parallel group sessions in the three groups. Wednesday Sept. 25 began with a joint session where each of the groups presented their work, which was then discussed jointly. The afternoon (until the excursion) was taken up by joint discussion on report structure. The morning of Thursday Sept. 26 encompassed joint discussion on a proposed joint manifesto as well as group discussions on application domains and application scenarios (topic areas were Health, SmartCities, Industry 4.0). The afternoon was taken up by continued group discussions and ended with group presentations and joint discussion on application domains and application scenarios. There was also further discussion on report structure at the end of the day. The manifesto was subject to very lively discussion in the evening, after dinner. Friday Sept. 27, the last day of the seminar, was

[2] https://www.internationaldataspaces.org/

devoted to wrap-up (conclusions, summary, and report process) followed by joint discussion on relations between Systems, Data and Business views on the overall topic of the seminar.

The outcome of the seminar, which is documented in the remainder of this report, encompasses summaries of the group discussions and the joint manifesto.

2 Table of Contents

3 Overview of Invited Plenary Talks

3.1 Recent developments of a legal framework for IT

Gerald Spindler (University of Göttingen, DE)

In his highly stimulating and provocative keynote address, law professor and high-level EU advisor Gerald Spindler shared important observations about a serious misfit between the conceptualizations pursued in business administration and computer science, and the structuring of the law system. As a consequence, judges are often surprised by seemingly unpredictable and contradictory answers to their legal questions. Conversely, managers and engineers are confronted with a law system that seems to adapt extremely slowly to the rapid progress, and with for them very surprising interpretation of this changing reality. For example, the speaker surprised the audience with the statement, that no concept of data ownership exists in Europe, except for the right to ones own personal data in the GDPR regulation. In the discussion, all participants agreed that joint research is urgently needed to better match the conceptual world of law and ethics, with the technical, user, and business perspectives on data ecosystems.

3.2 A few thoughts about managing business models for platform-based data ecosystems

Frank Piller (RWTH Aachen University, DE & MIT Media Lab, US)

In his keynote, Frank Piller started from the observation that – in contrast to traditional product platforms e.g., in the automotive industry – the value of smart products does no longer lie in the product itself, but rather in the connections in its business ecosystem. The resulting network effects and multi-sided markets have already been intensely studied, most visibly by 2015 Economics Nobel Prize winner Jean Tirole. The strategic question then becomes whether a company wants to offer a platform or an app. Successful platform businesses show striking differences from traditional organizations e.g., in terms of value created vs. value captured per employee. Current empirical research at RWTH Aachen University is studying how these concepts are being transferred to networks of German small and medium enterprises, e.g., in the smart farming sector.

3.3 Semantic Data Interoperability

Maurizio Lenzerini (Sapienza University of Rome, IT)

In his tutorial, Maurizio Lenzerini gave an logic-based structuring of four different data interoperability architectures: data integration, data exchange, data warehouses/data lakes, and collaborative data sharing. He pointed out the central importance of formal mappings

with different schemas, and then went into some depth into description logic-based approaches for compile time tasks such as mapping discovery, analysis, and reasoning, but also into runtime tasks such as data exchange, update propagation, data quality, and mapping-based direct and reverse query rewriting.

3.4 International Data Spaces

Boris Otto (Fraunhofer ISST & TU Dortmund, DE)

In his overview talk on the Fraunhofer-led International Data Spaces (IDS) initiative, Boris Otto first presented some typical examples from European industries, and cited worries of a majority of European SMEs about trust and security, data sovereignty in terms of keeping control over shared data, and inconsistencies in not just data interoperability, but also process interoperability. He then reported a large-scale, multi-year requirements study about desirable properties of a solution, from which the IDS reference architecture as well as a conceptual information model, several advanced algorithms (e.g., for data usage control) and suitable governance mechanisms for so-called alliance-driven platforms are emerging.

4 Technology

4.1 Systems

Boris Düdder (Univerity of Copenhagen, DK) and Wolfgang Maaß (Saarland University, DFKI, DE) and Julian Schütte (Fraunhofer AISEC, DE)

4.1.1 Motivation

Data ecosystems operating valuable data assets need strong guarantees on data security while fostering openness, community, and value creation.

Data is becoming a significant asset for any organization [23]. A key challenge is related to technical architectures for managing and processing data. A dominant approach centralizes data and distributes processing results (ref). Alternatives are less popular but feasible, such as peer-to-peer architectures or other more distributed approaches (ref) in which either data or program (query) are transferred optimizing for connection capacities. Related to the underlying architecture are organizational principles and governance structures. For centralized architectures, principal-agent logic is applied, i.e., a principal (user, client, customer) sends data to an agent (server, service provider) who processes the data and sends back results. This organizational architecture is applicable if the key value lies in the software as such, but not in the data. Now that data become a value as such, a bilateral principal-agent, or technically a client-service architecture, does not necessarily fulfill the requirements of data economic systems.

Several technical elements are essential key components for a data ecosystem, as explained in the following subsections, such as security and encryption, data analytics, AI and semantic technologies and ontologies, multi-agent systems, peer-to-peer architectures and, last but not least, algorithmic correctness and correctness of implementations.

4.1.2 Scope and Requirements

The system boundaries are used to define the purpose of our data ecological system w.r.t. technologies and applications. Therefore, these system boundaries influence fundamental design decisions of the problem model design and are reflected in the solution design. The boundaries are defined along the dimensions of the organization and technical scope. Organizational boundaries are organized along with entities' roles, rights, obligations, and prohibitions. Such entities are key concepts such as contracts specifying obligations, data/data objects/data sets representing valuable curated data assets, and semantics/metadata/ontologies generating information out of pure data. The technical scope should support system design and modules, various pervasive architectures, i.e., central, tightly coupled, and loosely coupled, needs for adaptability [12] and security, as well as protocols, computation and programming models.

Data formats and protocols are necessary for expressing contracts, i.e., for enforcing policies for data usage and conditions or guards such as pricing or billing for example. A challenge in a distributed network to enforce contracts and usage constraints at a remote peer, which is controlled by another participant. Possible ways for enforcement are by establishing trust in the participants and their data handling, e.g., employing remote attestation [26], or by private function evaluations [21].

4.1.3 State of the Art

Traditionally data is not perceived as a valuable asset as such but as an intrinsic part of any software. Data populates databases as a carrier of facts about a domain. Physicists, chemists, astronomers, biomedicial researchers, and economists understand for a long time that data is a basic asset that requires processing and filtering for deriving knowledge (e.g., CERN LHC). Companies, such as Google, Amazon, Microsoft, Baidu, and Alibaba, adopted this understanding for extracting knowledge that can be used for predicting human behavior. The success of these attempts was transferred to production industries (Industrie 4.0 and cyber-physical systems) and even private life (smart city and smart home). Thus, data becomes a valuable asset as such, called a data product [62]. Taking a product-driven approach, system designs start with the data product and ask in which market, to whom, and for what price it can be sold, i.e., applying a *product logic*. This adopts an inside-out perspective by which a data product is the starting point, and the market is the target. From an inside-out perspective, prices for data products are determined by adopting a cost-driven approach. Taking an outside-in perspective, the customer of a data product is the starting point by asking the question of which value can be created on the customer side by leveraging data products. This generally applies a *service logic*.

Data ecosystems are characterized by the type of business and the type of community [79]. The terms 'knowledge market' and 'data market' are used as synonyms nowadays. The type of business is distinguished between commercial and non-commercial data ecosystems while the type of community distinguishes between closed on open. Closed data ecosystems are established under the umbrella of one juridical person. This could be, for instance, a corporation, an association, or an individual person. Non-commercial, closed data ecosystems are intraorganizational while non-commercial, open data ecosystems provide data products unlimited for free or with the option for donations, e.g., UC Irvine Machine Learning Repository. Viable business models for data ecosystems are closed, for instance, by membership or proprietary bi- or multi-lateral contracts. Commercial, open data ecosystems are organized as marketplace [24] that support trading and transactions [4].

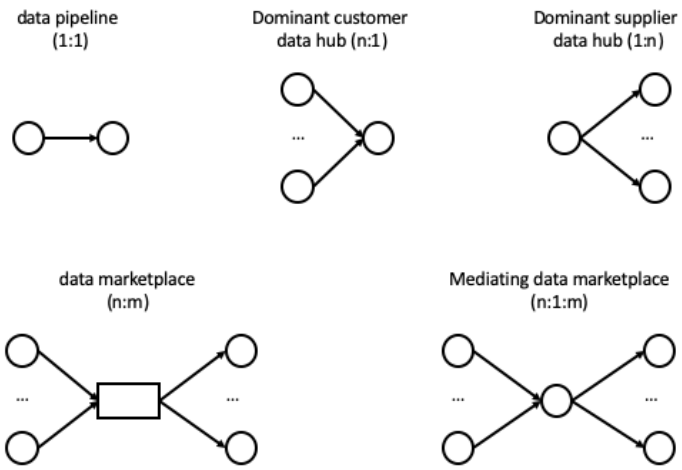


Figure 1 Basic system design patterns for data ecosystems.

System designs of data ecosystems taking a product logic adopt the concept of a marketplace in which the main interaction occurs by matching data product providers with customers. Facilitating roles are providing support services on financial transactions, logistics, quality assurance, insurance, and notary services. The main purpose of a product-driven marketplace lies in the transfer of access and usage rights under governance of explicit contracts or background regulations and laws.[3] The most simple system design is a *data pipeline* that establishes a marketplace between a single data provider and a single customer. Facilitating services are fully materialized by the marketplace and fixed contracts. More flexible are *data hubs* that allow $n : m$ transactions between data providers and customers. A $n : 1$ *data hub* is used if there is only one customer but many data providers. Dominant customers establish data hubs and invite data providers (*dominant customer data hub*). In contrast, a $1 : n$ *data hub* has only one or very few data providers but many customers (e.g., Airbus Skywise or Bloomberg Market Data Feed). A $1 : n$ data hub system design implements a standard product distribution network (*dominant supplier data hub*). A derivative of this system design is a *1:1:n data hub* with an independent agent implementing a data hub between one data provider and n customers. In reverse, the same system design can be applied with one dominant customer, i.e., $n : 1 : 1$ *data hub*.

More elaborated are $n : m$ *data hubs* or $n : 1 : m$ *data hubs* that constitute the concept of a *data marketplace*. Without dominant market players, a data marketplace is governed by offer and demand [80]. A data marketplace is considered being *liquid* if any demand can be contractually satisfied by an offer for a mutually acceptable price and low transaction costs. In plain data marketplace, trust between buyers and sellers is not guaranteed but requires additional trust-building regulations and additional roles.

The support for trusted and secure computation balancing participants requirements and expectations is challenging. Electronic marketplaces with multiple agents have been studied under various security objectives, including privacy, non-refutation, and accountability. The focus has primarily been on the digital transactions of such markets and how to secure them. The physical value-chain is aligned to the digital transactions of such marketplaces because

[3] According to Gerald Spindler's statements, data ownership cannot be transferred from a data originator to any other agent.

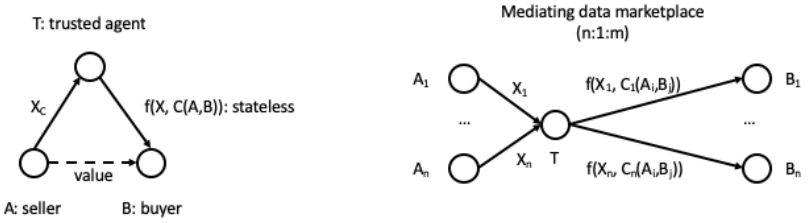


Figure 2 Generic system design for data marketplaces.

these marketplaces emphasize on physical goods as transaction value and not on data as a valuable good. To secure consistency and correctness of transactions, at least two different approaches are in use. The first approach employs monitoring and auditing of the processes, checking the consistency of good transfers and digital representation, as well as ensuring that data flows adhere to contractual agreements. The costly and laborious approach motivated due diligence [36] and various usage control frameworks [99, 92, 100]. Another approach is prevention by either using legal penalties to force compliant behavior of market participants or by technical measures including cybersecurity and cryptography. The latter has become a highly active area of research and several significant advances have been made in privacy-enhancing technologies such as verifiable computation and non-interactive zero-knowledge proofs [8, 1, 112, 18, 102, 7], distributed ledger technology [16, 17], and fully-homomorphic cryptography [50, 73]. Depending on the desired guarantees for the system, the correctness and robustness of algorithms and the system implementations are necessary. Security is not a product; it is a process, which demands a verified tool-chain, e.g., akin CompCert [77] and platform, i.e., trusted execution environments. The interaction of components and their reliability could be supported by automatic program generation [56, 11], DSLs [39] and verification methods.

4.1.4 Model

Data pipelines and data hubs replicate traditional interactions in business relationships between firms. Data marketplaces are a model class subsuming all other types of data ecosystems. Three roles are key to a data marketplace: buyer, seller, intermediary trusted agent. The role of a trusted agent is mandatory because of the transactional nature of a marketplace that requires a transaction being atomic. Data products X are provided from seller A to buyer B in exchange to another entity, in particular, financial assets. By doing this, value is created on the buyer side. In general, a seller A only transfers projection of X to buyer B for various reasons, such as IP protection and trade secrets. Nonetheless, a seller A might have a business incentive for selling certain properties and insights on X. Therefore, seller A and buyer B negotiate a function f that shall be allowed to apply on X under contractual requirements C. The role of the trusted agent T is that only f is applied to X as specified in C (Figure 2).

Expression $f(X, C(A, B)) : \textit{stateless}$ means that a function f is applied to a data product X or a set of data products according to specification given by contract C acting as a callback function in f, e.g., in [78]. The result is provided to buyer B according to specifications in contract C. Function f is required to be implemented in a stateless fashion, and no traces will be left with agent T. These needs to be assured by trust-building measures, such as program verification and certification. Therefore, a system design consisting of seller A,

buyer B, and a trusted agent T guarantees that a) data products are only used according to a specification, b) buyer B only gains access to resulting data, and c) no data traces are left with agent T.

4.1.5 High-level Description of Possible Applications

We present here two use case scenarios which could benefit from such a system.

Predictive maintenance – Predictive maintenance helps to determine the condition of operational equipment in-use for predicting when maintenance should be performed and which components are affected. Cost savings over routine or time-based preventive maintenance could be achieved because activities are triggered only when necessary. In this scenario, a production site creates raw data from production machine sensors, and the machine manufacturer is interested in running analytics on the sensor data to improve his customer maintenance service. On the other hand, the production plant owner is concerned about sharing raw data of the production machines, which could allow interested parties, e.g., competitors, to infer business secrets. The scenario can be enabled with a secure data sharing which balances the interests of both parties or even third parties.

Pharmaceutical companies – Pharmaceutical companies are producing and delivering drugs to patients through complex supply chains. The associated indirection between pharmaceutical company and patient complicates the due diligence, e.g., in the presence of adverse effects of drugs. Thus, sharing information between pharmaceutical companies, regulators, health practitioners, and patients could improve drug safety and reduce the costs of due diligence. All parties in this scenario have interests to share data but, at the same time, have high interests in restricting the data exchange, e.g., for privacy or competitive advantages. The system is realizable using secure data sharing allowing to create a data ecosystem harmonizing the needs of all participants.

4.1.6 Research Challenges

The project is ambitious, and to our best knowledge, there is no product on the market dealing with the identified challenges on business, legal, and computer science side. Business is founded on trust and trust-building activities, which is expensive to build and to maintain. The research challenges identified are (semi-)automatic contracts negotiation and agreement on functions as well as function execution. Semantic information of data and functions is necessary to automate negotiation and agreement on functions. On the other hand, data-/function-specific guarantees of market participants need to comply with data protection rules. The treatment of data in law is very often focused on privacy but not on data as an asset. The consequent juridical uncertainty, e.g., on the data ownership and copyright, is a business risk and poses a challenge for law researchers as well as legislation. To replace trust with mathematical and technological guarantees is an additional research challenge for computer science as a discipline. The verification of functions, their execution in secure execution environments, and secure and trusted libraries are not available on the market and require research from various interdisciplinary and inter-subdisciplinary research. The system challenges are compelling but indispensable for the future success of data ecosystems.

4.1.7 Conclusion

Because of the interdisciplinary aspects of data ecosystem platforms, computer science can provide various models, methods, and tools to support the development of such platforms.

The challenges are not only located in computer science but also the interplay with other disciplines. The challenges and opportunities presented in this section are novel in combination with their usage domain and its inherent restrictions, e.g., legal or economic. We postulate a platform and sound formal models supporting the activities described in this report and offering solutions to the more general problems of data ecosystems.

4.2 Data

Cinzia Cappiello (Polytechnic University of Milan), Yuri Demchenko (University of Amsterdam, NL), Ugo de'Liguoro (University of Turin, IT), Bernadette Farias Lóscio (Federal University of Pernambuco, BR), Avigdor Gal (Technion – Haifa, IL), Sandra Geisler (Fraunhofer FIT – Sankt Augustin, DE), Maurizio Lenzerini(Sapienza University of Rome, IT), Paolo Missier (Newcastle University, GB), Barbara Pernici (Polytechnic University of Milan, IT), Jacob Rehof (TU Dortmund, DE), Simon Scerri (Fraunhofer IAIS – Sankt Augustin, DE), Maria-Esther Vidal (TIB – Hannover, DE)

4.2.1 Scope

In this section, the main outcomes of the group of *Data* are reported. First, diverse data-driven problems are presented; solutions to these problems demand to effectively describe and integrate heterogeneous data, accurately represent and assess data quality, and ensure data specifications during the execution of operators or queries. Next, the state of the art is summarized and the proposed model for data ecosystems is defined. Finally, insights of the research challenges to be addressed are outlined.

4.2.2 Motivation and Requirements

Data-driven technologies in conjunction with smart infrastructures for management and analytics, increasingly offer huge opportunities for improving quality of life and industrial competitiveness. However, data has grown exponentially in the last decades as a result of the advances in the technologies for data generation and ingestion. Moreover, data is usually ingested in myriad unstructured formats and may suffer reduced quality due to biases, ambiguities, and noise. Thus, the development of efficient data management methods are demanded for enabling the transformation of disparate data into knowledge from which actions can be taken in scientific and industrial domains. Problems to be addressed include:

- Definition of a generic data market architecture that can describe data properties such as economic goods and data exchange models. In this type of architectures, data lakes play a relevant role in the storage of huge amount of heterogeneous data on distributed infrastructures.
- Active data networks described in terms of specifications and collections of components which can be repositories of data or queries. Data sources are connected using schema mappings. The evaluation of queries against the network requires the composition of data sources respecting the restrictions imposed by the connections existing the components of the network.

- Accurate description of meta-data of data sources that can change over time and effective usage of these descriptions whenever data is shared.
- Models that estimate the cost of integrating different sources, and the benefits that the fusion of new data sources adds to the accuracy of query processing.
- Hybrid approaches that combine computational methods with human knowledge; limitations and benefits of these approaches in the resolution of data-driven problems, e.g., schema matching, data curation, and data integration need to be established.
- Paradigms for making data suitable for sharing are required; they should ensure trustworthiness, e.g., in terms of data quality or data market players.
- Federated query processing for addressing data interoperability issues during query execution demands meta-data to select the relevant sources for a query. Furthermore, data quality assessment is required for determining the quality of the answers produced during query execution against a set of selected sources.
- Management of new generated data from data analysis and machine learning performed on existing sources.

4.2.3 State of the Art

4.2.4 Data Integration

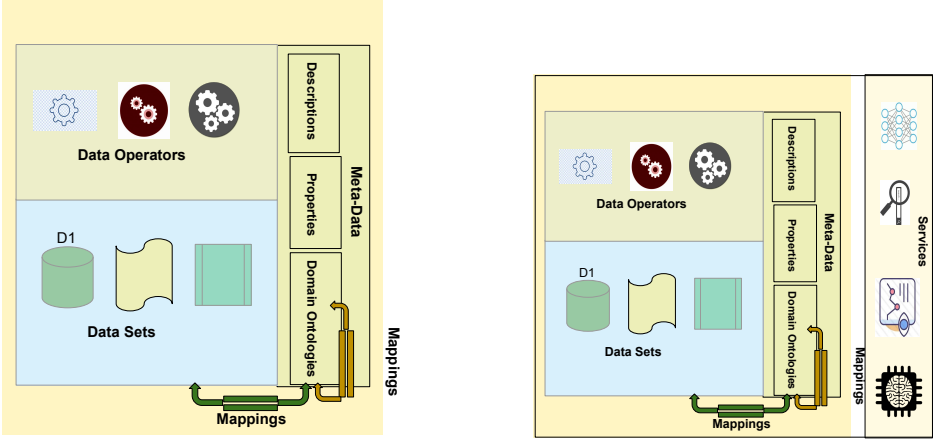
The problem of integrating data collected from different data sources has been extensively treated in the literature [34, 55]. The mediator and wrapper architecture proposed by Wiederhold [114] and the data integration system approach presented by Lenzerini [74], represent the basis for the state of the art [27, 52, 65, 83]. Following these approaches a global ontology encodes domain knowledge and enables the description of the meaning of the data to be integrated by means of mapping rules. Since creating mappings manually in a data integrating system is a tedious and time consuming task, diverse approaches have been proposed to discover mappings in a (semi-)automated way (e.g., CLIO[42], IncMap [94], GeRoMe [61, 54], KARMA [65]). Additionally, a vast amount of research has been conducted to propose effective and efficient approaches for ontology alignment or schema matches [22, 41, 45, 81, 85]; modeling and management of uncertainty of the schema matching process is the paramount importance for providing a quantifiable analysis of the accuracy of the discovered mappings [44].

4.2.5 Data Ecosystems

Data ecosystems are a special kind of digital ecosystem. As such they are distributed, open and adaptive systems with the characteristics of being self-organizing, scalable and sustainable. While being centered on data, the main concern with data ecosystems is about knowledge sharing and growing, which is at the same time an issue of learning from unstructured and heterogeneous data, construction of new abstractions and mappings, offer of services, including querying, data integration and transformation. All this should be ensured in a dynamic and scalable way, while retaining consistency, quality assessment, security and affordability.

4.2.6 Query Processing Over Heterogeneous Data Sets

Existing solutions to the problem of query processing over heterogeneous data sets rely on a unified interface for overcoming interoperability issues, usually based on metamodels [63]. A few Data Lake systems have been proposed, mainly with focus on data ingestion and metadata extraction and management. Exemplary approaches include GEMMS [96],



(a) A Basic Model for a Data Ecosystem. **(b)** A Data Ecosystem equipped with a brain.

Figure 3 A Basic Model for a Data Ecosystem.

PolyWeb [64], BigDAWG [37], Ontario [40], and Constance [54]. These systems collect meta-data about the main characteristics of the heterogeneous data sets in a Data Lake, e.g., formats and query capabilities; additionally, they resort to a global ontology to describe contextual information and relationships among data sets. Rich descriptions of the properties and capabilities of the data have shown to be crucial for enabling these systems to effectively perform query processing.

4.2.7 A Model for Data Ecosystems

A data ecosystem *DE* is defined as a 4-triple *DE=<Data Sets, Data Operators, Meta-Data, Mappings>*; Figure 3a depicts a data ecosystem in terms of its components.

- *Data sets*: the ecosystem is composed of a set of data sets. Data sets can be structured or unstructured; also, they have different formats, e.g., CSV, JSON or tabular relations, and can be managed using different management systems.
- *Data Operators*: the set of operators that can be executed against the data sets.
- *Meta-Data*: provides the description of domain of knowledge, i.e., the meaning of the data stored in the data sets of the data ecosystem. It comprises:
  - i) *Domain ontology* provides a unified view of the concepts, relationships, and constraints of the domain of knowledge. It associates formal elements from the domain ontology to each *D*. For instance, *workshop* and *participant* can be part of the concepts in a domain ontology.
  - ii) *Properties* enable the definition of data quality, provenance, and data access regulations of the data in the ecosystem. For instance, *last updated* and other non-domain properties (quality etc).
  - iii) *Descriptions* of the main characteristics of a data set. No specific formal language or vocabulary is required; in fact, a data set could be described using natural language. For instance, *Data set D is about a Dagstuhl seminar.*
- *Mappings* expressing correspondences among the different components of a data ecosystem. The mappings are as follows:

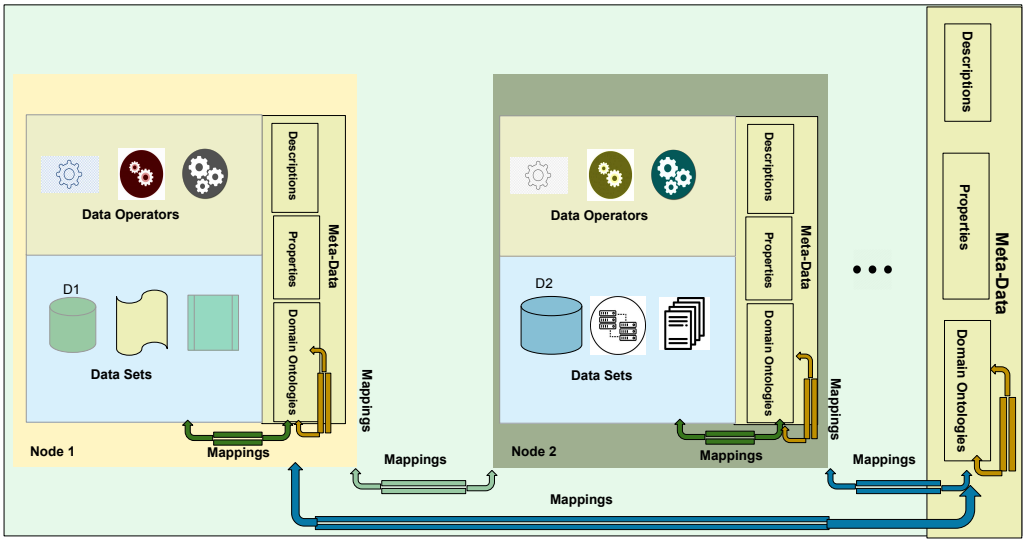


Figure 4 A Model for a Network of Data Ecosystems.

- *Mappings between ontologies*: they represent associations between the concepts in the different ontologies that compose the domain ontology of the ecosystem.
- *Mappings between the data sets*: they represent relations among the data in the data sets of the ecosystem and the domain ontology.

A data ecosystem can be equipped with a "brain", able to execute services against the data sets (Figure 3b). Services include query processing, data transformation, anonymization, data quality assessment, or mapping generation. The services are able to exploit the knowledge encoded in the meta-data and operators to satisfy the requirements of the applications implemented across the ecosystem. The following correspond to examples of services:

- *Concept discovery*: identify a new concept, e.g., *foreign student*, using machine learning. Based on the result, the domain ontology and the mappings can be augmented.
- *Data set curation*: a service able to keep humans in the loop can be used to create a curated version of a data set in the ecosystem. The service can also update the properties of the ecosystem to indicate the provenance of the new curated data set and manage new generated data from data transformation, analysis, and learning.
- *Procedure synthesis*: a service able to construct new procedures out of elementary build blocks by composing existing services toward new goals. In a complex and evolving system, it would be unfeasible to program procedures and even queries without automatic support; also, the exploration of repositories and libraries of existing procedures should be available.

A set of data ecosystems can be connected in a network. For this, we envision an *ecosystem-wide* meta-data layer where the entire ecosystem is described. Figure 4 depicts a network where nodes and edges correspond to data ecosystems and mappings among them, respectively. In this configuration, the meta-data layer describes each of the nodes in terms of descriptions, properties, and domain ontologies. The following types of mappings can be defined among the nodes of a network of data ecosystems:

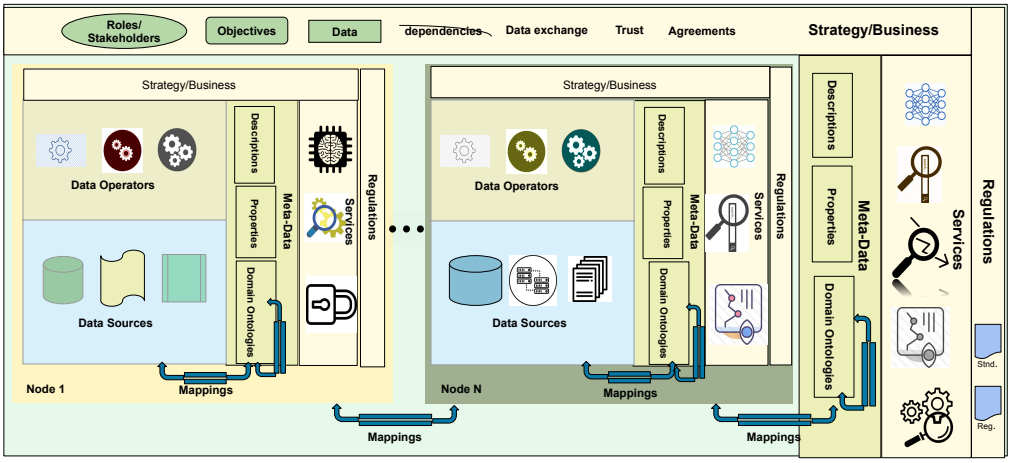


Figure 5 A Model for a Network of Data Ecosystems Empowered with Strategy and Business Models and Regulations.

- *Mappings between domain ontologies*: they state correspondences among the domain ontologies of two nodes or between one node and the global meta-data layer. For example, the concepts *workshop* and *seminar* in nodes N_1 and N_2, respectively, are the same.
- *Mappings between properties*: they describe relationships among properties in two nodes. For example, the provenance of two curated versions of a data set could be the same.
- *Mappings between data sets*: they represent correspondences between the data sets in two nodes. For instance, mapping data from D_1 in node N_1 to D_2 in node N_2, can represent that the list of students in a university 1 is the same to list of students in university 2.

Finally, data ecosystems can be enhanced with additional meta-data layers to enable the description of business strategies and the access regulations. Figure 5 depicts the main components of a network of data ecosystems empowered with these layers. As can be observed, this enriched version of a network of data ecosystems comprises:

1. Meta-data describing business strategies will enable the definition of the stakeholders of the network and their roles.
2. Objectives to be met and the dependencies among the tasks that need to be performed to achieve these objectives.
3. Agreements for data exchange and criteria for trustworthiness.
4. Regulations for data access and for data privacy preservation.
5. Services composing services of the nodes of the network or exploiting their operators.

4.2.8 Research Challenges

This section presents an outlook on the main challenges to be addressed in the implementation a network of data ecosystems.

4.2.9 Data and Metadata Curation

The availability and quality of data resources must be ensured, so that data value creation can be stimulated. A promising solution is to use a well-conceived, efficient curation strategy for data resources and their metadata. Such a curation technique is the continuous process of

managing, improving, and enhancing the data and their metadata. Furthermore, the curation process aims to ensure that the data and metadata meet a defined set of quality requirements, such as security rules, integrity constraints, or metadata availability expectations. Without proper curation, data resources may deteriorate in terms of their quality and integrity over time. One of the major challenges for achieving continuous curation of metadata is to create a methodology to structure the curation process as well as to provide a set of tools. Furthermore, data sets obtained through transformation processes of data analytics or machine learning can become new data sets. Not only provenance but also the processing methods should be associated with them.

4.2.10 Data Traceability and Data Consumption Monitoring

Tracking the utilization of data sets and applications using these data resources is still a big challenge. Such information could be useful for both the identification of new data sets and for data quality improvement. Monitoring data consumption, as well as providing effective ways for the consumer to interact with the data publisher, should make it possible to collect information about using and sharing data. In this sense, it is crucial to obtain consumers' feedback in such a structured way that it allows identifying flaws in the published data, the need to publish new data, and, for example, to enable classification of data.

4.2.11 Describing the Main Properties of Data Sets

All data sets in a data ecosystem should be described in terms of their main characteristics. In particular, in the case of numerical data, a general framework that allows for the representation of uncertainty and imprecision is required. In addition, keeping track of data set updates and modifications, and allowing basic access to versioning information are also important issues. For instance, a new version may be created when there is a change in the structure, contents, or characteristics of a data set. As data sets can change over time, maintaining different versions of the same data set and enable access to them becomes necessary.

4.2.12 Data Enrichment via a Joint Human/Machine Integration Effort

Data integration in a data ecosystem is challenged by the need to handle large volumes of data, arriving at high velocity from a variety of sources, which demonstrate varying levels of veracity. Data integration has been historically defined as a semi-automated task in which correspondences are generated by matching algorithms and subsequently validated by a single human expert. The reason for that is the inherent assumption that *humans do it better* which is not necessarily the case. A current challenge involves the identification of respective roles of humans and machines in achieving cognitive tasks in a way that maximizes the quality of the integrated outcome. We observe that the traditional roles of humans and machines are subject to change due to the availability of data and advances in machine learning.

4.2.13 Adaptive/Context-aware Data Quality Assessment and Redefinition of Data Quality Dimensions

Data quality assessment requires the selection of both the dimensions to be evaluated and the metrics to measure the selected dimensions. An accurate evaluation of the quality of the data depends on diverse factors, e.g., the type of source or data, and the application that aims to use the data. This implies an adaptive approach for data quality assessment able to trigger the appropriate metric. Furthermore, data quality assessment should be performed

in two phases: the registration and the usage phase. During the phase in which data are registered/ingested into a data ecosystem, a first evaluation is needed. This evaluation should consider some basic metrics in order to guarantee that low quality data is not ingested in a data ecosystem. On the other hand, in the usage phase, metrics that reflect the suitability of a data set along the applications that aim to access the data set need to be defined. Moreover, data quality can be measured at different levels. There are quality dimensions which refer to the schema or overall structure of a data source. Additionally, data quality dimensions regarding the content of the data sources can be defined. In both cases, metrics which represent functions to determine the corresponding values for the data quality dimensions, may cover atomic items, such as a single attribute, or multiple items, or complex objects. In a single data ecosystem node, but especially in the case of sharing data between several nodes, it is not clear, how data quality values should change, when the data is further processed. The way to calculate new values for a single dimension might be dependent on the meaning of the dimension and a given metric, as well as specifically on what happens to the data in the current processing step.

4.2.14 Measuring Information Gain in a Data Ecosystem

Adding new data sets to a data ecosystem arguably results in some kind of information gain. The gain extends naturally the increase in knowledge that occurs when new mappings are discovered within a data ecosystem node – either interactively or semi-automatically, or across diverse nodes in a data ecosystem. Questions that need to be addressed include:

i) How is the information gain defined and measured?
ii) Can the gain be defined in terms of benchmark queries and the results they return given an evolving state of a data ecosystem?

4.2.15 Exploring the Social Dimension of a Data Ecosystem

Users (humans, or data consuming services) are actors in a data ecosystem, playing an important and active role in its evolution. Firstly, they can provide feedback on data sets that they have used (either formally or informally), adding to the quality and provenance properties, which other users should be able to take into account. Secondly, they may decide to curate data sets that exhibit poor quality properties and return a better version of those to the DE, generating more metadata in the process (provenance tracing with details of the interventions). Finally, they may contribute to the mapping exercises that the DE makes possible and indeed relies upon. Some of the questions that emerge when users are first class citizens include:

i) Should data ecosystem users form a network, and if so, can such a network be used to recommend / promote / deprecate data sets?
ii) Should users (or services) be credited for producing and making available new (better) versions of a data set, i.e., through curation activities?
iii) How can the social layer of the DE be promoted, maintained (is that a part of the "upper layer"?), and exploited?
iv) How would a recommender system be able to suggest data sets to users based on their usage history of the data ecosystem and the implicit connections with other users?

4.2.16 Explainability in Data Ecosystems

Whenever the *brain* of a data ecosystem executes a service (e.g., query, data transformation, quality evaluation, or mapping generation), it should be able to explain the result of the operation carried out in executing the service using an appropriate language. There is a connection between provenance and explanation. In some sense, provenance information can be part of/exploited in an explanation. Included in the notion of explanation is the idea of explaining the semantics of a data source–either primitive, or produced by a data operation.

4.2.17 Query Processing over Data Ecosystems

Effectiveness and efficiency of query processing over a federation of data sources is known to be affected not only by the number of sources in the federation but also for the heterogeneity of these sources. In case of a Date Ecosystem, data sources can be ingested in different formats and accessible with management systems that provide different data management capabilities, e.g., some systems may support complex queries including joins and aggregations while others do not. Moreover, the data quality and structuredness may considerably vary. For ensuring the correct processing of queries, interoperability issues should be solved at different levels. Meta-data about the description of the meaning of the data stored in the data ecosystem plays a crucial role for enabling the execution of queries over the data sources that will provide the correct answers. Moreover, if several versions of the data sets are stored in a data ecosystem, the process of data integration must take into account the temporal dimension of data. For example, answering queries like *tell me all the students of University X who have become professors at X after no more than 10 years from graduation*, might involve looking at several versions of different data sets.

4.2.18 Data Ecosystems for enabling the specification of Meta-Metadata

The proposed model for Data Ecosystems enables the definition of a second layer of global meta-data which describes both data and meta-data descriptions of all the local nodes in the Ecosystem. Managing and keeping meta-metadata *up-to-date* is more challenging because both data and metadata may change as a consequence of the dynamics of such systems. Dynamicity may be at least twofold:

i) local dynamics of node data and metadata and
ii) global dynamics of queries and mappings involving several nodes and the Data Ecosystem metadata themselves.

The research questions to answer under these conditions are:

a) How to define safety criteria?
b) What kind of transformations should be allowed?
c) Should we prevent non-conservative structural updates?
d) Which level of data integrity and consistency will be ensured?

4.2.19 Specification language for procedure synthesis

Beside ontologies and meta-data, a language for specifying services and their functional properties is needed. This is not just for the sake of verification, but mainly because they are required for procedure synthesis. For automatic search through large libraries and for synthesis to be feasible such a language should balance expressive power with effectiveness and efficiency. Attempts to use type theory for these purpose include work by Bessai et al [10] and Hengelin and Rehof [58].

4.3 The Business of Data Ecosystems

Elda Paja (IT University of Copenhagen), Matthias Jarke (RWTH Aachen & Fraunhofer FIT), Boris Otto (Fraunhofer ISST & TU Dortmund) and Frank Piller (RWTH Aachen)

4.3.1 Background and Motivation

▶ **Example 1** (Data Ecosystem). *Consider the example of PRINT INC., an industrial print shop reacting to increasing demand by adding more flexible production capacity. The firm provides machine vendor DRUCKMASCHINEN AG a list of requirements, derived from its most sophisticated print jobs. DRUCKMASCHINEN, however, lacks insights about the specific situation and suggests a standard specification. After ramp-up, a sample analysis reveals an OEE far below the anticipated values, as configurations, parameters and procedures of the existing facilities do not recognize the capabilities of the new equipment. This scenario is still a very common situation in many industries. With a data ecosystem in place, PRINT INC. could first provide DRUCKMASCHINEN better past operating data to better determine requirements. A specialized analytics service provider would assist in this analysis, also drawing on technology insights and usage data from other print shops operating in a related setup. After implementation, a higher level of OEE could be reached by learning from best practices from similar operations of other manufacturers. By getting access to rich usage data – covering not just machine data, but also data on the production context, material characteristics and behavior of PRINT INC – OEE can be continuously improved. The data ecosystem enables a continuous feedback of this data into the development cycle of DRUCKMASCHINEN, revealing requirements for the next generation of production hardware and software. In times of low capacity utilization, PRINT INC. provides DRUCKMASCHINEN access to its new printing machine, which becomes part of DRUCKMASCHINEN's production network, connecting printing capacities from hundreds of printers virtually. With its PRINT HUB platform, DRUCKMASCHINEN can place print jobs from clients like publishing houses (without own production capacity) on existing machinery, generating a new business opportunity for both itself and its customer. At the same time, by moving from the role of a manufacturer and service provider of printing machine to that of an operator, DRUCKMASCHINEN also gains critical knowledge in the operating principles and continuous improvement opportunities of its machinery.*

To realize this typical vision for Industry 4.0, much more than a technical infrastructure is required. Beyond the availability of data and capacity, this scenario only can be realized if it has been determined how the value gained by accessing these inputs is captured and shared among the actors involved:

- Would profiting from such external data also require to provide similar feedback data?
- Could PRINT INC. control who and what is printing on its capacity shared via PRINT HUB?
- How could print shops differentiate themselves by process know-how when a data ecosystem enables a kind of instant benchmarking, as all operational data is shared over the network?
- Why is DRUCKMASCHINEN, in the end, not operating all its machines directly, integrating vertically into the domain of its former clients?
- Who owns the governance of this network? Is it DRUCKMASCHINEN as the focal keystone player, or should rather an alliance of print shops, users (like publishing houses) become the owner of the platform [113]?

This generic example provides a glimpse of the challenges from an economic and competitive dynamics perspective that come with the vision of a data ecosystem, which is, by definition, not restricted to a focal company or value creation within a closed network of established partners. It resembles the vision of an open network of sensors, assets, products, and actors that continuously generate data. This data is utilized to enhance operational efficiency, but also to provide new opportunities for strategic differentiation. A core element hence is a business-model perspective on **(re-)usage of data, insights, and applications by other parties** than those generating the data at the first place. Generally, learning and analytics can take place faster and more efficiently if manufacturers not only utilize their own data, but also could access data from similar contexts in other industries.

Actual research on industrial data-based business models has focused predominantly on the perspective of value creation, i.e. how to use shared data to create new service offerings like predictive maintenance, energy optimization, quality improvements, etc. [19]. But the open question still is an understanding how **to incentivize the sharing of deep production know-how and data** in order **to balance value creation (using the data) with value capture (sharing the rents)** [70]. The **rise of platforms** where these data are being exchanged and enhanced by dedicated "apps", often offered by specialized third-party entities, is **one of the largest economic developments of the last decade** [28, 91]. As platform **interfaces** become more open, more actors will be attracted into a data ecosystem, and the platform orchestrator will be able to access a larger set of potentially complementary innovative and operative capabilities. Most of the external contributors will innovate in ways complementary to the platform. Some, however, may start developing **capabilities** in ways that become competitive to the platform. Such an emergence of competition will depend on the **governance of the ecosystem**. Collaborative governance mechanisms will increase complementors' incentives to innovate in platform-enhancing ways. This demands especially a dedicated setup to share the value created on the original platform, but is also dependent from the technical interface design. In turn, emergence of competition from former complementors is likely to create a reaction by the platform leader to start competing back with these new rivals, either by enveloping them, or by closing its technological interface, in effect **moving away from being an open data ecosystem** towards becoming a more closely managed network or internal platform. Based on a discussion of different governance frameworks for data sharing and access, this workgroup of the **Dagstuhl Seminar #19391** discussed various scenarios, but especially questions for further research. Considering actual developments in platform-based business models that recently emerged also in manufacturing industries, our workgroup discussion tried to understand the design of the business models, but especially how to incentivize (reward) the sharing of deep production know-how and data in order to balance value creation (using the data) with value capture. This led to the following questions suggested for future research:

- Modeling the tension between openness in value creation and control of value capture, while recognizing the need for establishing and increasing trust in data sharing.
- Managing property rights (access, transfer, enforcement) at data, applications, and connected assets as a result of varying degrees of platform openness
- Definition of governance modes and design factors to generate adequate business models for a data ecosystem that allow to maximize value appropriation for all involved actors

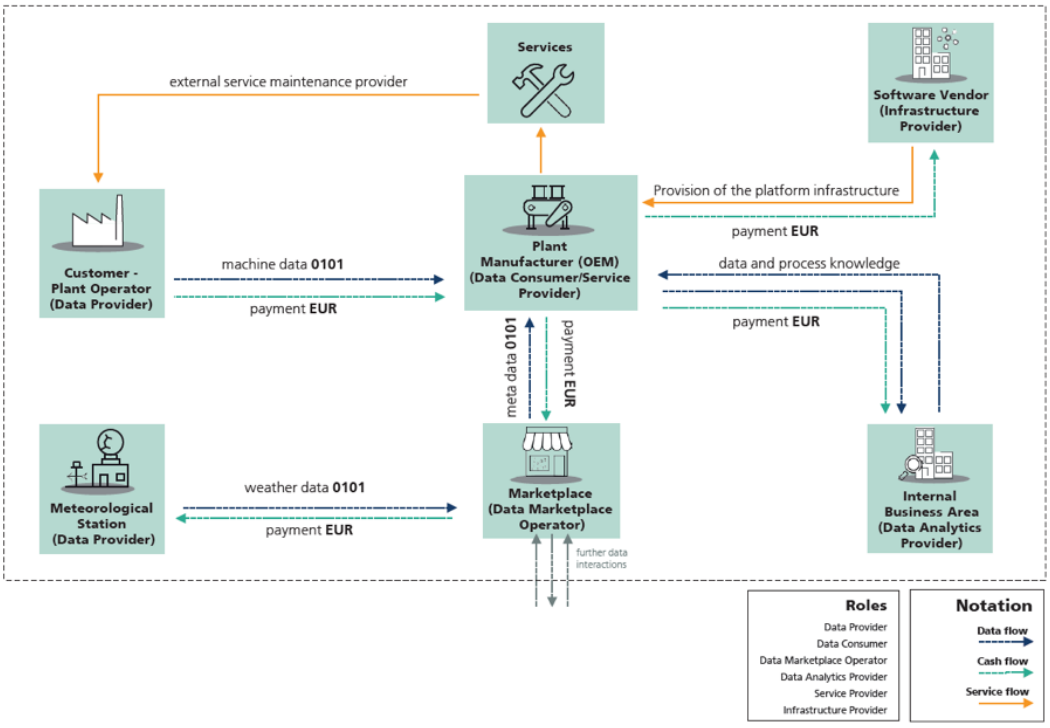


Figure 6 Digital Business Ecosystem Example.

4.3.2 From Data Ecosystems to Business Ecosystems

Figure 6 illustrates the interactions within a digital business ecosystem in terms of data, cash and payment flow. Within the ecosystem perspective, a holistic view on the business relations is provided.

The digital ecosystem is enabled via multiple technical infrastructures. In this example two major infrastructure providers (data marketplace operator and software vendors provide a network to connect various entities such as the physical machines, customers, service staff, and third party services e.g. for data analyses. The Original Equipment Manufacturer (OEM) collects data through specially installed edge devices on the customer's sites. The operational machine data is transferred to the platform-based cloud infrastructure provided by the OEM. The data is stored, processed, and analyzed on the OEM's platform. Part of the data analysis is performed by the OEM itself within separate organizational units, so that the role of the data analytics provider is also taken internally. Therefore, additional historical and product development data is provided from internal resources to improve the analysis. Based on the analyses of the OEM, various services can be provided to the plant operators via the platform (e.g., dashboards on machine usage, specific operational reports, process optimization recommendations, maintenance planning). New external data sources (e.g., weather data) from meteorological stations can be considered as new parameters for analysis. This data is provided within a marketplace, which the OEM uses to enrich the existing data basis. The data marketplace acts as a gateway to other data ecosystems and promotes cross-industry data exchange. The ecosystem perspective provides an overview of the dynamics within a practical data ecosystem including the flow of data, payment and services.

4.3.3 State of the Art

One of the most important questions of a firm is how to efficiently create value: Should it produce its own output or should it orchestrate the output of others? In the case of software, the choice increasingly favors orchestration over production. Apple, Google, and Microsoft became the three most valuable companies in the world in 2015 by relying on a **platform business model** to provide software applications for its users, developed by independent app programmers. An open platform business model offers distinct economic advantages because it allows a firm to harness external inputs and innovation as a complement to internal innovation [91]. These platform markets are, however, neither a new phenomenon nor restricted to software or information goods. Open platforms have emerged in aerospace (Lockheed Martin), T-shirts (Threadless), 3D printing (MakerBot), and shoes (Adidas).

Since the early 2000s, the industrial organization literature has begun to develop theory on platforms (also referred to as "two-sided markets", "multi-sided markets", or "multi-sided platforms" [97, 98]). Economists view platforms as special kinds of markets that play the role of facilitators of exchange between different types of users that could not otherwise transact with each other. Essential to most economic definitions of **multi-sided platforms** (MSPs) are the existence of "**network effects**" that arise between the "two sides" of the market [49]. As the value of the platform stems principally from the access of one side to the other side of the platform, the question of platform adoption becomes how to bring multiple sides on board. Such platforms typically reside upon a **layered digital infrastructure**, where lower-level layers (e.g., physical components, transmission layer) enable and support functionalities at higher, user-facing layers (e.g., operating systems layer, application layer) [109, 115]. To create value, ecosystems hence depend on complementary inputs made by loosely interconnected, yet independent stakeholders from varying levels of (technological) distance from the end consumer [3, 90].

MSPs[4] facilitate the establishment of business ecosystems, which are formed by the users interacting on the platform. Researchers have examined a number of case studies on MSPs in which a keystone firm owns and governs the platform [86, 104]; so these are "keystone-driven MSPs". However, the platform landscape is becoming more and more diverse, with other, more complex governance and ownership structures being observable in different domains. De Reuver et al. [30] have found that in many cases there is no single platform provider, but that the platform is jointly designed and "shaped" by multiple actors. In this context, Tiwana et al. [110] point to the importance of distinguishing platforms owned by a single firm from platforms characterized by some form of "shared ownership". Shared ownership materializes in multiple organizational forms, among them the alliance. Gawer [48], for example, identified some MSPs from the supply chain management domain that are "shared among firms that are part of a formal alliance". Such "alliance-driven MSPs", which are characterized by shared ownership and governance (e.g. a joint-venture company or industry association), as well as decentral platform governance models have been neglected by academic research so far [30].

Table 1 compares keystone-driven with alliance-driven MSP design.

In the case of an **industrial data ecosystem**, platform participants include the **orchestrator of the platform** (today either IT infrastructure providers like SAP or Software AG, or providers of automation or manufacturing equipment like Siemens, GE, Trumpf), operators of **production assets** (users in form of "factories"), which provide data, providers

[4] This paragraph is taken from Otto and Jarke (2019) [89]

Table 1 Juxtaposition of Keystone-Driven and Alliance-Driven MSP Design.

Theoretical Concept	Keystone-Driven	Alliance-Driven
Platform architecture	Architecture determined by goals of keystone firm	Architecture determined by shared interest of multiple owners (leading to decentral data storage, for example)
Platform boundary resource	Mainly technical boundary resources (APIs, SDKs etc.), supported by "social" boundary resources (e.g. training for developers) [14]	Data (IDS specific) as a boundary resource of "dual" nature, i.e. requiring both technical processing and functional use; many social boundary resources, such as working groups, task forces etc
Platform design	Core developed by platform owner, then extended by complementors	Consensus oriented design process with focus on "common denominator"
Platform ecosystem	1) Innovation, 2) Adoption, 3) Scaling [57]	1) Adoption, 2) Innovation, 3) Scaling
Ecosystem governance	Start with limited number of sides and limited options for interaction between them, then increase number of sides and options for interaction [105]	Start with complex ecosystem (i.e. multi-stakeholder setting), then reduce to core ecosystem and extend it later on depending on roll-out requirements
Regulatory instruments	Mainly pricing instruments, accompanied by non-pricing instruments	Dominated by non-pricing instruments; integration of pricing instruments scheduled for scaling phase; data governance

of applications analyzing this data and providing prescriptions and predictions (**app programmers**), and other asset operators that utilize insights from aggregated data to optimize their own production. In addition, also the goods being produced can become part of the platform in form of **connected ("smart") products**, providing a feedback loop of usage data, but also becoming the center of another platform around digital services complementing these products. The latter also refer to **end-users (customers)** as a final participant of the ecosystem.

An applied stream of research looked into the demand for business model innovation (BMI) in leu of Industrie 4.0 [19, 69]. Based on rather qualitative research approach, a demand for professionalizing the BMI process in established companies was identified [95]. Similarly, specific aspects of BMI for Industrie 4.0, i.e. specific patterns or components of I40 BMs, have been identified. By analyzing Industrie 4.0 characteristics for a firm's business model and the approaches of pioneering companies to instigate BMI, this research developed a methodology to both generate, model, systematically design, and evaluate BM alternatives triggered by Industrie 4.0 [93].

Another recent stream of literature complements the original economic analysis of platforms with a more managerial and design-orientated perspective. It looks into the distinct **governance and orchestration challenges** presented by established innovation ecosystems (e.g., [33, 113]). This literature has looked, for example, into prices, incentives, contracts, and network effects. Less work has addressed how prospective ecosystem stakeholders **commit resources** towards a de novo ecosystem creation effort and how they evolve a shared structure of interactions [28]. Yet, this is exactly the situation of many data ecosystems, where creation is not a simple endeavor of an app. Rather, the ecosystem value proposition depends on the concurrent availability of complementary inputs from varied, independent stakeholders.

In such a situation, a core decision of a platform operator is about how much **to open the platform** and when to absorb inputs (developments, apps, data) from the connected parties. This decision drives adoption and harness developers as an extension of the orchestrator's own production function [9, 15, 91]. [28] extend the analysis to contexts in which the establishment of a new business ecosystem cannot be reliably planned, as no clear value proposition to orchestrate the ecosystem exists ex-ante. In a piece of complementary research, [9] investigate the perspective of platform complementors (app programmers) and their decision to join a platform based on its openness. The authors develop and validate a platform **openness measurement instrument** that captures perceived platform openness.

Since 2014, our Fraunhofer Industrial Data Space initiative [60, 87] has focused on requirements and rather technological challenges of inter-organizational data exchange. This requires novel conceptual information modeling at multiple levels, and still significant research for the specialization to the case of production engineering. The ideas pursued at the boundary of CRD_A and this workstream build on significant prior research on the conceptual modeling, structure and evolution analysis, and data-based Social Network Analytics concerning strategic dependencies and trust among players [46, 47, 84].

4.3.4 Conceptual Background[5]

The proliferation of digital technologies and AI accelerates a shift in business models which is characterized by an increasing importance of data as a strategic resource. While traditional business models rest on tangible assets, data is the raw material not only for information and knowledge, but for innovative services and customer experiences. Besides the shift from tangible to "smart" products and from controlling the physical to orchestrating the data value chain, there is one additional fundamental change in the digitalized economy. Innovation increasingly takes place in ecosystems in which various members such as businesses, research organizations, intermediaries such as electronic marketplaces, governmental agencies, customers, and competitors band together to jointly achieve innovative value offerings.

Ecosystems are characterized by the fact that no one member is capable of creating innovation on its own, but the ecosystem as a whole needs to team up. In other words: Every individual member has to contribute in order to benefit. Ecosystems function in an equilibrium state of mutual benefits for all members.

In a data ecosystem, data is the strategic resource for the success of the whole system as it is understood as a stand-alone asset that will be exchanged and monetarized within the ecosystem. That offers the participating actors new growth opportunities through networking with other participants and acts as a driver for innovative services and customer experience.

Data sharing opens up new opportunities for progress and the formation of cooperations with other companies or actors from which every participant in the data ecosystem benefits. The various activities of the different members in a data ecosystem lead to a complete coverage of the data value chain. This includes the stages of data generation, curation, exchange, storage and analysis as well as the use of the resulting knowledge for comprehensive business decisions. Through the sustainable data exchange the participating actors are able to develop further and to operate value co-creation which leads to new digital value proposition.

[5] Cf. Otto et al. (forthcoming). [88]

4.3.5 Data as a Boundary Resource[6]

As far as the boundary resources are concerned, Henfridsson and Bygstad [57] argue that this concept is helpful for studying patterns of interaction between the various groups and agents on a digital platform. Boundary resources are resources through which different agents create relationships and interact with each other in order to co-create value [38]. Dal Bianco et al. [14] distinguish between technical and social platform boundary resources. Typical boundary resources are Application Programming Interfaces (APIs) and Software Development Kits (SDKs). Examples for social boundary resources are intellectual property rights and documentation of software services. Furthermore, boundary resources are not stable, but evolve over time. Eaton et al. [38] coin the notion of "distributed tuning" to describe the process of continuous shaping and reshaping of boundary resources between the different platform actors and users. More recent research has suggested to increasingly look at such boundary resources of digital platforms as a promising subject of analysis [30].

How organizations can exchange and share data has long since been an important research topic. The need for companies to exchange and share data has been a major motivation for the development of platforms mediating between suppliers and buyers of goods. Early two-sided data exchange solutions were facilitated by technological standards, such as EDIFACT or ANSI X.12. Gawer [49] within her integrative platform framework identified traditional buyer-supplier relationships for which data is a technical enabler.

Around the turn of the millennium, electronic marketplaces emerged as intermediaries to reduce the complexity of the increasing need of n:m data exchange [101], in which data from multiple sources (n) can be bundled and utilized in contextualized presentations to multiple users (m). This intermediary function comprised – among other things – the mapping of the different message schemas of the various standardization initiatives that evolved. Motivated by the success of peer-to-peer-networks in the consumer realm, some researchers explored technologies, and even business models, for peer-to-peer based networks for data exchange in the industrial domain. Technological aspects of peer-to-peer data ecosystems, such as context exchange among different world views of organizations [51] or automation of data mappings in heterogeneous settings [61], have been investigated since the late 1990s. In 2005, Franklin et al. [43] noted the growing richness of digital media and proposed that users should be enabled to create their own "data space", where a free collection of data and media objects could be managed under a user specific network of semantic metadata. In 2010, the notion of the "data lake" was coined [72] and quickly received attention in the practitioners' community. Furthermore, some researchers investigated the role of data within platform based ecosystems [59, 68, 82]. More recent research has dealt with the upcoming phenomenon of data platforms, mainly encouraged by the discourse around big data [13, 31, 59]. These studies focus mainly on platform architecture technology and data flows.

4.3.6 Research needs

While the emergence of data ecosystems offers new opportunities for the different ecosystem participants, many social, environmental and business challenges have to be addressed in order to pave the way for these opportunities to materialize. Among the most significant challenges are:

[6] Section taken from Otto and Jarke 2019 [89]

- Trust: New methods are needed to increase trust in data sharing so that more data would be available for new applications. What is needed is a framework that includes building blocks for data sharing, data management, data protection techniques, privacy-preserving data processing and distributed accountability and traceability. In addition to providing technology for platform developers, the framework should provide incentive and threat modelling tools for data sharing business developers and strategists, who consider opening data for new cooperation and business.
- Data Sovereignty: The framework should also support compatibility with the latest and emerging legislation, like the EU's General Data Protection Regulation (GDPR) and free flow of non-personal data, as well as ethical principles, like IEEE Ethically Aligned Design. This will increase trust in industrial and personal data platforms, which will enable larger data markets combining currently isolated data silos and increase the number of data providers and users in the markets. The result should aim to be platform-agnostic to be applied in multiple domains with platforms based on different technologies.
- Interoperability: The main objective should be to support a trusted data ecosystem providing easy-to-use privacy mechanisms and solutions that guarantee citizens and business entities can fully manage data sharing and privacy. The challenge is thus to provide a corresponding overall technical architecture which needs to take into account the key reference platforms and technologies to support data sharing, to improve existing solutions and architectures, to define the overall reference architecture, and to design platform-agnostic trusted data sharing building blocks and interoperability.
- Data Governance: Data Ecosystems highly depend on access to data and interactions of actors providing or using data or similar resources such as application programming interfaces (APIs). The role of data governance in these complex networks between organizations is an under-researched field. There is a lack of concepts and mechanisms to mandate responsibilities among participants of a data ecosystem. It is essential to study inter-organizational mechanisms that allow participative interactions, incentives to influence the dynamics and evolution of the ecosystem.
- Compliance with Antitrust Legislation: To avoid the risk of data monopolies, the following needs to be ensured:
  - Improving the mobility of non-personal data across borders in the single market, which is limited today in many member states by localization restrictions or legal uncertainty in the market;
  - Ensuring that the powers of competent authorities to request and receive access to data for regulatory control purposes, such as for inspection and audit, remain unaffected;
  - Making it easier for professional users of data storage or other processing services to switch service providers and to port data, while not creating an excessive burden on service providers or distorting the market.
- Data Economics: Data business, i.e. viewing data as an economic asset, will bring additional motivation for data providers and owners to open up their data for various applications. Personal data is becoming a new economic asset class, a valuable resource for the 21st century that will touch all aspects of society. The rapid development of the Personal Data Service (PDS) market will provide big changes in the way individuals, business and organizations deal with each other, as individuals assert more control over their data or service providers process personal data.

4.3.7 Data Infrastructures

Definition. *A data infrastructure is a distributed technical infrastructure consisting of components and services which support data access, storage, exchange, sharing and use according to defined rules.*[7]

The **International Data Spaces (IDS)**[8] initiative aims at data sovereignty for businesses and citizens in Europe and beyond. The IDS Association (IDSA) provides a reference architecture that enables an ecosystem for the sovereign exchange of data with clearly defined usage rights. The reference architecture defines a technical infrastructure and includes contractual regulations: at the semantic level, data linking, or analysis can technically be prevented or made possible. In this way, the classic structure of cloud services is also embedded in an interoperable digital economy with full data sovereignty on the digital infrastructures of third parties. The IDS standard solves a market obstacle: In order for data to unfold its value creation potential, it must be described and tradable according to a global and interoperable standard. This has never existed before. But DIN SPEC 27070 (to be published in Nov 19) is the first global and interoperable standard. 100 member institutions from the EU as well as from Brazil, Canada, China, India, Japan and the USA are involved; they come from all branches of industry and have developed information and governance models in the IDSA as the basis for the IDS architecture and its data sovereignty standard. More than 50 concrete application scenarios and first products are now available from companies of all sizes and sectors – together they are working on operational concepts for a sovereign data exchange infrastructure. The certification scheme "IDS_ready" also enables companies outside the association to participate in secure, IDS-based value-added processes via certified participants and components. Reference implementations and sample codes are available for developers and can be tested in testbeds. IDSA is in continuous coordination with global initiatives (Industrial Internet Consortium, OPC Foundation, Robot Revolution Initiative, BDVA) and participates in EU research projects to anchor IDS architecture and data sovereignty standards within European digitization strategies. In 8 countries, there are contractually bound IDSA hubs that bring standardization and adaptation of the technology to the respective country.

4.3.8 Questions for further research

From this brief review of literature in the field, but especially our conclusions during the seminar, we draw three conclusions:

1. The basic mechanisms of platform markets are well understood, especially the basic effect of network effects and complementing value creation in multi-sided markets.
2. Platform openness has derived as a key variable in the systematic design of a platform ecosystem. Literature suggests some factors that constitute openness that can be utilized for the design of a data ecosystem.
3. All existing analysis, however, has been conducted in the field of either consumer electronics or information industries (gaming, search engines, social media sites). Dedicated research in the context of industrial data applications is missing.

[7] For the source of this definition, please see https://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.pdf?___blob=publicationFile&v=18 (in German, English version to follow very soon).

[8] For the source of this section, please see https://www.internationaldataspaces.org/wp-content/uploads/2019/10/IDSA-digital-summit-international-statements.pdf

Future research in the field is required in various aspects, which can be structured in four layers of analysis, as suggested by Gawer [49]:

- **Interfaces**: On a technical level, the openness of APIs and other technical interfaces is not just a question of programming and quality control, but first an important design factor for the ability of connected asset to perform predictive and prescriptive functionality, i.e. to enhance its capability in this regard by getting access to data also from other actors. From the perspective of a platform, the openness of an API is a signal of willingness to share data and knowledge, hence potentially attracting third parties. At the same time, open interfaces are not just a technical risk, but also reduce the ability of the originator of the data to capture unique value from this data and hence differentiate it from other market players. Research needs to investigate the choices made by companies on different levels of a platform to understand the decisions and their consequences with regard to interface design. This also includes the deployment of Software Development Toolkits (SDK) by a platform operator, which determine the ability of third party application providers.
- **Capabilities**: This layer deals with capabilities that organization need to acquire to position themselves in an data ecosystem and corresponding platforms. These capabilities include business model innovation, mastering organizational change, or capabilities of orchestrating a manufacturing ecosystem. This research builds on a rich literature of capability building and organizational sense making and will study how dedicated capabilities link to firm performance. Of particular interest are question of counterbalancing (the lack of) capabilities of one actor by capabilities of another actor on a platform. This leads to a re-interpretation of the central economic question of the boundaries of a firm.
- **Organizational design** refers to the design of a platform ecosystem and the design factors ("business model patterns") that allow for value creation and capture in these industrial data platform. A particular focus of this research will be on the level of value capture, i.e., on mechanisms that allow the different actors of a platform for profit from their participation and contributions of the platform. Building on the last work package, this also asks the question whether firms shall joining an existing ecosystem (and of, which one and under which conditions) or try to orchestrate an own?
- **Governance modes**: This final layer integrates the previous work on platform-based value creation and value capture from an external, but also an internal perspective. A central construct in this work is the degree of openness vs. desire for control of each actor in the ecosystem. Future research needs to identify possible governance modes (and patterns of platform governance) and match those to the performance of observed use cases. This should help us to understand the choices made by managers in setting up these governance modes and will (for example, game-theoretical) model the theoretical consequences of the choices under given contingencies.

4.3.9 Outlook: How to compete when all become the same (have access to the same data)

The use of networked and intelligent production systems and dynamic value creation networks in the context of a data ecosystem accelerates a faster exchange of "best practices" across corporate boundaries. For instance, this can be attempts to the process optimization in production networks, but also access to the same complementary service of a platform provider. Thus, operational efficiency can be increased for an entire industry. However, competitive advantages do not result from operational efficiency, but from strategic uniqueness! As a

result, existing business models will be challenged. Companies that are highly focused on operational efficiency are facing increasing competitive pressure. Efficiency gains cannot be narrowed down to just one company. Instead of being entirely focused on operational efficiency, business models for data ecosystem must therefore initiate new differentiation opportunities for companies. The options for a more efficient and scalable custom performance have already been addressed. Openness and transparency can also become a differentiating feature: Not only during the innovation process, companies need to act more open instead of cutting themselves off. This also includes an intelligent and transparent handling of data. This intelligent handling of data enables the development of differentiating potentials using customer-specific knowledge. For example, this can be used to generate additional services or new products that satisfy the customer's benefit even better. Transparency and fairness can thereby become a competitive advantage. For companies, the challenge is to reach and maintain the competitive position as leading innovator in a specific industry. This can only happen with the best possible mix of open innovation processes and internal innovation. Companies need to find an implicit or explicit trade-off with regard to the openness of innovation processes and the type of shared knowledge. This includes finding strategies of how to implement innovations faster to the companies without internal barriers causing significant delays.

5 Use Cases

5.1 Use Cases from the Medical Domain

Sandra Geisler (Fraunhofer FIT – Sankt Augustin, DE), Maria-Esther Vidal(TIB – Hannover, DE), Elda Paja (IT Univerity of Copenhagen), Maurizio Lenzerini(Sapienza University of Rome, IT), Paolo Missier (Newcastle University, GB)

In the health care domain a variety of use cases exist, where data ecosystems are beneficial and open up new opportunities via data exchange. In the following, we describe two use cases which build complex data ecosystems in detail and analyze the challenges which are posed by them and other use cases in the health domain to data ecosystems.

5.1.1 Use Case: Multi-Site Clinical Tríal

In this example application for a data ecosystem, in the course of the research project SALUS (Selbsttonometrie und Datentransfer bei Glaukompatienten zur Verbesserung der Versorgungssituation)[9] funded by the Federal Ministry of Health, a multi-site clinical trial with glaucoma patients is conducted over one year in Germany where about 2000 patients will be included in the study starting in 2020. Glaucoma is a chronic disease of the eye with various causes possibly leading to irreversible damages of the eye nerves. For glaucoma patients it is crucial to keep the intra-ocular pressure in certain bounds to not exacerbate the condition which may lead to blindness in the worst case. The usual diagnostics require the regular creation of a status quo of the patient's condition, including an intra-ocular

[9] https://www.ukm.de/index.php?id=innovationsfondsprojektsalus

pressure profile over two successive days. For this procedure the patient has to be admitted to hospital. In the SALUS trial the advantages and disadvantages of using a mobile device at home to create a pressure profile over one week, called self-tonometry, is compared to the method applied in hospital. Additionally, the patient is equipped with a 24-hour blood pressure device, and at examinations questionnaires will be completed by the patient.

5.1.2 Study Process

Patients are included in the study by a local ophthalmologist who will explain the study to the patient and assign her to either the self-tonometry group or a control group randomly. Successively, the patient will be sent to a local hospital which will do the screening examinations and a study nurse will train the patient in using the self-tonometer. After the examination, one week of self-tonometry or a two day hospital stay, respectively, follow. The current therapy is adapted by the local ophthalmologist based on the results, if necessary. At regular intervals follow-up examinations are executed by the local ophthalmologist. After 12 months a final examination is done in the local hospital. After the trial, the collected study data will be evaluated by a research institute combining it with additional data provided by the health insurance companies of the patients.

5.1.3 Data Processing

During and after the trial all collected examination data, device data, and images of each patient have to be available especially to the local ophthalmologist, but also to the local hospital, treating this particular patient. They need the data to analyze the current condition and adapt the therapy where necessary. Hence, a glaucoma health record is maintained for each patient at a central server. Data from the self-tonometry and the 24-hour blood pressure monitoring device is transferred via a mobile device to the corresponding device manufacturer's portal. From there it is transferred anonymized to the trial server. The data and images from the examinations by the ophthalmologist and hospitals are also transferred from their systems and devices via a web application in an anonymized way to the central server. After completion of the trial, the data from the trial server has to be integrated with data of the health insurance companies and transferred to enable a comprehensive analysis by a research institute.

An overview of the use case, the involved nodes and stakeholders, and the corresponding dependencies aligned with the DE model from Section 4.2 are presented in Figure 7.

5.1.4 Challenges

We identified multiple challenges for this use case, which may also be generalized to other data ecosystems in the health domain. By the time writing, the project is still in an early stage, but where applicable, we sketch strategies for tackling the challenges.

Data views and access control. For each stakeholder, different views on the data have to be provided, as every role in this complex scenario has different access rights to the data. Hence, a very fine-granular definition of access control rights has to be provided to respond to this requirement. The data from the local hospitals, local ophthalmologists, and the device manufacturer services has to be anonymized and integrated to provide the data for the different views mentioned at different points in time during and after the trial.

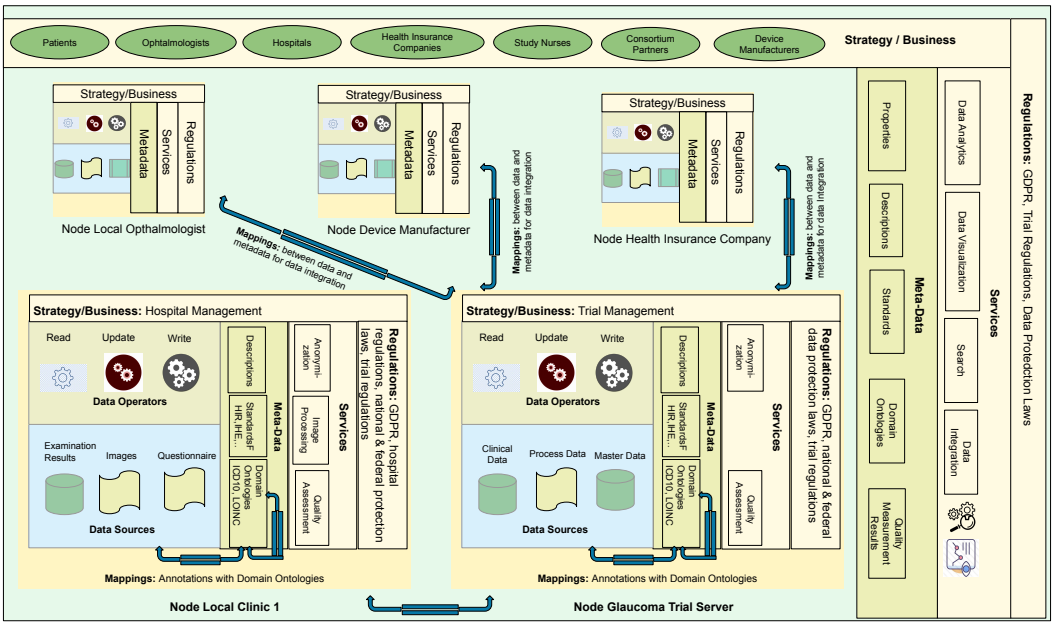


Figure 7 Overview of the Multi-Site Clinical Trial.

Data Integration. Thus, we need an integrated patient-centric global schema for a glaucoma health record where all data from each patient identified by a global identifier is combined. It has to be considered, if this integration should be virtual or materialized, and, how the mappings between the global schema and the sources can be created. In this use case most likely materialized integration will be used as the main data sources will not change frequently. Further integration of external data, such as medication information or usage of standards for enrichment and documentation, is a challenge given that suitable sources have to be found, their quality has to be checked, and the corresponding mappings have to be created. In this scenario various heterogeneous data sets and standards are involved which makes the integration a challenge. A data lake as basic data management architecture for the health record could be a solution to integrate various heterogeneous data sources and also allow for complex meta-data management, data quality management, access control, and search on the data. For each party/view a separate data mart could be created as a subset of the data, strictly controlling the access to the data. But such an architecture would imply a high implementation overhead as mature data lake systems are still not the usual case. A more simpler solution to be flexible according to the data storage could be also a single NoSQL database, where meta-data management and data quality management have to be realized separately.

Consent management vis-a-vis the GDPR. A further challenge is the implementation of consent management and usage control in compliance with the GDPR. Corresponding means for the deletion or editing of data at one or multiple available sites have to be implemented if the patient revokes the consent or asks for corrections. Furthermore, the GDPR also demands for transparency of operations on personal data, which requires a form of provenance tracking and auditing. Data protection and secure communication also have to be ensured, when data is exchanged between the nodes. Finally, quality monitoring during the trial can be implemented to get high value data.

5.1.5 Use Case: Precision Medicine and Health Policy Making.

In this case study, the focus is on the integration of structured and unstructured data ingested from clinical records, medical images, scientific publications, or genomic analysis. The application of a network of data ecosystems is illustrated in the context of the EU H2020 funded project iASiS [10] which aims at exploiting Big data for paving the way for accurate diagnostics and personalized treatments. iASiS[11] is a 36-month H2020-RIA project that has run from April 2017 to March 2020, with the vision of turning clinical and pharmacogenomics big data into actionable knowledge for personalized medicine and decision making. iASiS aims at integrating heterogeneous Big data sources into the iASiS knowledge graph. As input of the problem, we have a set of myriad sources of knowledge about the condition of a lung cancer patient. Electronic health records (EHRs) preserve the knowledge about the conditions of a patient that need to be considered in order to have effective diagnoses and treatment prescriptions. Albeit informative, EHRs usually preserve patient information in an unstructured way, e.g., textual notes, images, or genome sequencing. Furthermore, EHRs may include incomplete and ambiguous statements about the whole medical history of a patient. As a consequence, knowledge extraction techniques are required to mine and curate relevant information for an integral analysis of a patient, e.g., age, gender, life habits, genotypes, diagnostics, treatments, and family medical history. In addition to evaluating information in EHRs, physicians rely upon their experience or available sources of knowledge to identify potential adverse outcomes, e.g., drug interactions, side-effects or drug resistance. Diverse repositories and databases make available relevant knowledge for the complete description of a patient's condition and the potential outcome. Nevertheless, sources are autonomous and utilize diverse formats that range from unstructured scientific publications in PubMed[12] to repositories of structured data about cancer-related mutations. Various services need to be implemented in order to transform relevant data that come in different formats into a common format and for data anonymization. Additionally, since the same concept can be identified with different identifiers, the detection and representation of mappings between data sets is necessary. Each data provider establishes a regulation about the type of operators and services that can be executed over each data set; they can also indicate which of the users of the data ecosystem can execute the operators and the services. Finally, GDPR regulations need to be respected ensuring that personal data is used according to the consents provided by patients. Figure 8 illustrates an overview of a solution precision medicine use case using a network of data ecosystems.

5.1.6 Challenges

Clinical data is usually stored in diverse formats, e.g., in notes in Clinical Records, gene sequencing panels, or medical images. Additionally, these data sources may suffer from potential biases, ambiguities, and noise. To overcome these data issues, distinct knowledge extraction methods need to be included as part of the services of the network. Typical extractions methods include:

i) *Natural language processing (NLP)*;
ii) *Visual analysis and image processing*; and *Genomic Analysis.*

[10] http://project-iasis.eu/
[11] http://project-iasis.eu/
[12] https://www.ncbi.nlm.nih.gov/pubmed/

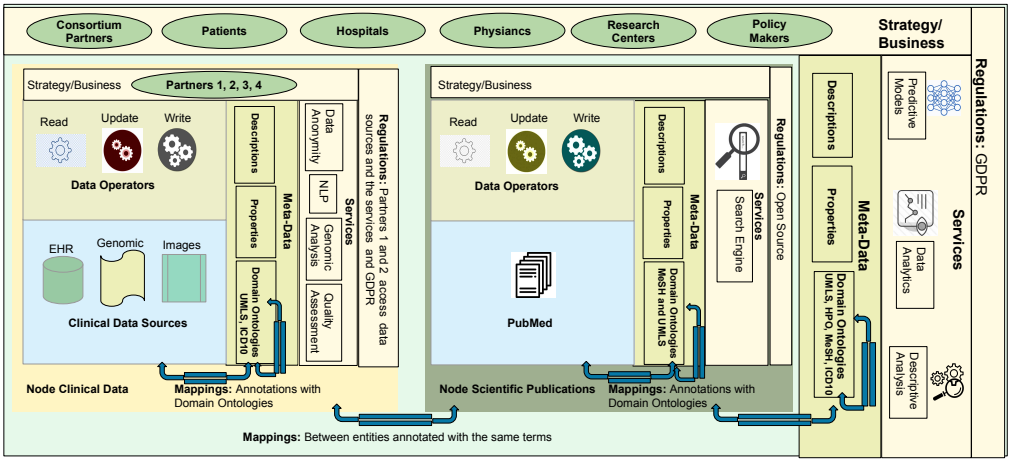


Figure 8 Overview of a Biomedical Data Ecosystem. A Network of Data Ecosystems for the Precision Medicine Use Case; it comprises two nodes of clinical data and scientific publications.

Additionally, data quality assessment techniques are required in order to identify data quality issues, as well as to define strategies for data curation. Data integration taking into account the meaning of the biomedical concepts is also challenging as the consequence of the variety of formats and representations. Moreover, exploring and visualizing data ecosystems require the implementation of scalable methods able to traverse a variety of data sources. Finally, privacy and data access control techniques are demanded to enforce regulations imposed by data providers or data protection authorities. The project aims at developing data management techniques that enable for the transformation of unstructured into a unified knowledge base. Domain ontologies like UMLS, are used to annotate the extracted concepts; these annotations provide the basis for entity matching and data integration. A unified schema is utilized to describe the integrated entities and a federated query engine supports the exploration of the knowledge base. Moreover, services for data analytics and prediction allow for the discovery of novel patterns and associations that may explain the survival time and disease progression. Initial results suggest that these knowledge discovery techniques on top of a knowledge base have the power of uncovering relevant patterns in the integrated clinical data. [71, 111].

5.1.7 General Challenges in Health Applications

5.1.8 Multitude of Stakeholders and Patient-centric Data Exchange

In the health domain many different stakeholders with diverse interests, rights, and responsibilities exist. In contrast to other domains, not only organizations are involved in use cases for DE, but also single persons, such as patients and physicians. In many cases personal data is involved which poses many challenges to the data ecosystem especially in terms of data access and data control. The role of the data owner may be taken by a party (e.g., the patient) different from the one who controls access to the data (e.g., the physician or clinic). This opens up many questions as data owner and "data controller" need to clearly negotiate which (transitive) rights the data controller may have and rules on how data exchange and usage is handled need to be defined. Though the handling of personal data is quite strictly and clearly regulated by laws, such as the GDPR, it is not an easy task to define a framework which suits all possible use cases and stakeholders' needs. A patient-centric data ecosystem

would need to enable transparency of the data exchanged and of the operations executed on them, making data provenance and auditing come to the fore. A patient-centric DE should give a patient the opportunity to control what is exchanged with whom and with what constraints. Finally, it should convey trust, that data is handled correctly to the rules and constraints negotiated in the DE. Solutions to implement access and usage control for DE have already been proposed, e.g., in the International Data Space [87]. These differ in various aspects, such as the invasiveness for the internal systems (which is a crucial point for health domain use cases) or the policy languages they are using.

5.1.9 Data Anonymization vs. Data Usefulness

A further concern is the trade-off between possible anonymization of personal data and its usefulness. Depending on the use case, anonymization of the data is a must, e.g., in clinical trials and the analysis of the results, or it is desired by the providing party to not disclose certain data. But the anonymization may reduce the usefulness of the data for further analysis. Hence, it is interesting to analyze the trade-offs between the extent of anonymization and the resulting usefulness. These can be used to check against the requirements of an application intending to use the data. Usefulness and anonymity could be criteria to search for suitable data sources in a data ecosystem.

5.1.10 Data Variety, Standards, and Interoperability

In the health domain, a huge variety of data formats is used, many of which do not apply recognized standards, but rather are proprietary to specific devices or systems. Although a multitude of standards exist in the health domain, these are not used consistently and for many crucial concepts, such as Electronic Health Records, a general standard is still missing. In these cases a DE has to demand a precise documentation and meta data description of the data provided to make the data interoperable for other parties. In the first (professional health services) and second health market (consumer health and well-being services) many insulated applications and data silos exist. Applications are just written for a specific purpose or device, but are not able to connect to other systems or even export into a standard format. To make this data reusable in other contexts, other services need to encapsulate it or wrap it. Hence, a DE should provide means to easily integrate these data sources into a DE, e.g., enable services, which transform the data into an accepted standard or (semi)-automatically add semantic annotations. For these services extra costs for data consumers or providers could be charged.

5.1.11 Meta-data Usage, Quality, and Mappings

Many medical taxonomies and ontologies exist to express meta-data and to annotate data. This is beneficial on the one hand to make the data more “understandable” and enables more intelligent applications, which also consider the context and meaning of the data. On the other hand, many of the most used vocabularies are so huge, such as MeSH, SNOMED, or the NHI Thesaurus, that working with them and maintaining them is very difficult. This results in quality issues in the vocabularies and, as a consequence, also in the applications that are based on them or the mappings created between two or more ontologies. Also data curation as such is an error-prone and tedious process, leading to poor annotations when this is performed by untrained persons or by algorithms.

5.1.12 Data Quality

Data quality (DQ) plays a crucial role in all data-intensive applications, hence, also in medical applications. There exist several sources of DQ problems in health applications. Much of the health data is recorded or transformed using electronic devices and sensors which may have an inherent physical imprecision or are easily prone to failure and lead to DQ problems. The monitoring of DQ for online or streaming sources is difficult and may lead to additional problems, such as synchronization and timeliness problems. Additionally, a lot of data is collected manually, e.g., using paper forms, which are also prone to errors, missing data, or transfer errors when digitizing the data. Furthermore, a multitude of data sources and formats may be involved in just one application. The conversion between formats, the integration with other data into existing data sets or databases, and the aggregation of data may as well lead to DQ problems. Many standards, for example FHIR, provide guidance to structure the data accordingly, but it can be very different how the guidance is implemented which may also lead to problems in the data format and interpretation. Also, the semantic consistency of data (is the data representing the patient and her condition correctly?) of single and between multiple examinations and measurements must be checked which often has to be a manual task. It is interesting to investigate how integration of several sources and Machine Learning algorithms could be used to support this process.

5.1.13 Data Protection and Data Sovereignty

Working with personal data is highly regulated by laws on different levels. For example, considering Europe and the EU, data security and protection is regulated by the General Data Protection Law (GDPR). Each country in the EU additionally may have a national law (e.g., in Germany the Federal Data Protection Law), which extends the directly applicable GDPR. Further regulations can be defined on state level (e.g., the Krankenhausgesetz in Bavaria or the Data protection Law in North-Rhine Westphalia). These multiple overlapping regulations lead to a highly complex environment for data ecosystems and data exchange. In turn, this requires a highly flexible framework which must combine usage control, access control, data provenance, and further means to enforce data protection and enable data sovereignty.

Data sovereignty for patients is desirable, but not easy to implement. Patients may not feel capable of deciding which information should be accessible to whom. It is hard to decide for laymen, what exactly is sensitive data and if this data is crucial to a specific health professional to enable the best possible treatment. The grade of transparency regarding data exchange may also be subject to discussion, as for example a very "chatty" notification protocol may annoy or make patients feel not secure.

5.1.14 Conclusion

In this section we have presented two use cases which constitute complex DE in different ways, but sharing similar problems. In general, the challenges for data exchange are manifold and specific in the health domain as many stakeholders with many different interests are involved. Especially, single persons, such as patients and doctors, play a major role in the DEs which requires to overthink mechanisms so far identified and implemented for industry and companies. Highly sensitive data and the corresponding ethics and laws around it, pose crucial challenges to DEs. And finally, the high variety in data formats and metadata definitions make the domain description and application implementation special. Hence, we postulate that a framework for DEs has to take all of these challenges and requirements for the health domain into account to propose general solutions.

5.2 Industrie 4.0 Data Ecosystems Examples

Egbert-Jan Sol (TNO – Eindhoven, NL)

Multi-sided markets are based upon an alliance agreement for sovereign data exchange between organizations and result in a distributed data driven ecosystem. In such data ecosystems (DES) each organization can do more then when using only its own data, and once properly architectured this can be realized without the need to see all data such that each party remains control over its own business.

Industrial data ecosystems are encountered in the (discrete) manufacturing, the (chemical) process industry, in (maintenance) and logistics services, but also in amongst other industries as telecommunication ecosystems as GSM. In practice all these industrial cases there is a physical device that can get a so-called digital twin, being the digital representation of a physical good. With this digital twin more data can be linked to the device resulting in data sets that can be (partly) shared. In industry we encounter data sets that consist of sensor data, copyrighted data as drawings all the way up to very valuable, confidential data sets.

We introduce in this section four notions: 1-a data analogy of material, 2-digital twins and identifiers, 3-four classes of data and finally 4- a function of controlled data exchange. It is followed by four industrial use cases of data ecosystem: GSM mobile telephony, discrete manufacturing, traceability of food/goods, and preventive maintenance with hints to aspects as use of an association, international standard and cyber security protection consequences.

5.2.1 From data element to data sets with sovereignty control to value

Data records, due to their digital nature, are often seen as discrete elements. But to accept that certain data is more valuable than other data, an analogy with the material/process industry might be more applicable.

Say, an individual sensor reading could be seen as atom and a long set of sensor readings as base material. Sensor data as such is not so valuable and sensor data is not copyright protected. Then if one combines iron with carbon to produce steel or mix a polymer with a color ingredient one can get a stronger alloy due to the iron/carbon matrix/lattice or a colored plastic. Combine two data set also leads to a more valuable set then just a single list of data points, just as steel is more valuable than its base materials. Next image the combination of a drawing and a set of manufactured parts that are verified to be produced from the steel alloy or plastic within specs according to the drawing. This set is more valuable than just a bag of polymer granules or a block of steel. Data sets that are verified are also more valuable.

We won't identify the single sensor data element as we won't identify an individual atom. The base material is already more valuable, but not as valuable as a strong alloy. Similarly we might identify a list of sensor points having some value, but the combination with a safe boundary set within the sensor points must stay is already more valuable. And an end product similar as a list of customer bills is even more valuable.

There is an analogy in data set similar to atoms, to base materials to chemicals, to discrete products up till a product owned by someone. Single data lists, data sets and their relations and finally a whole data bases can be have a certain value for a user. And combining data sets over multiple parties and share it in a larger data ecosystem can even be more valuable. The challenge is to construct the data ecosystem such that parties can do more with the

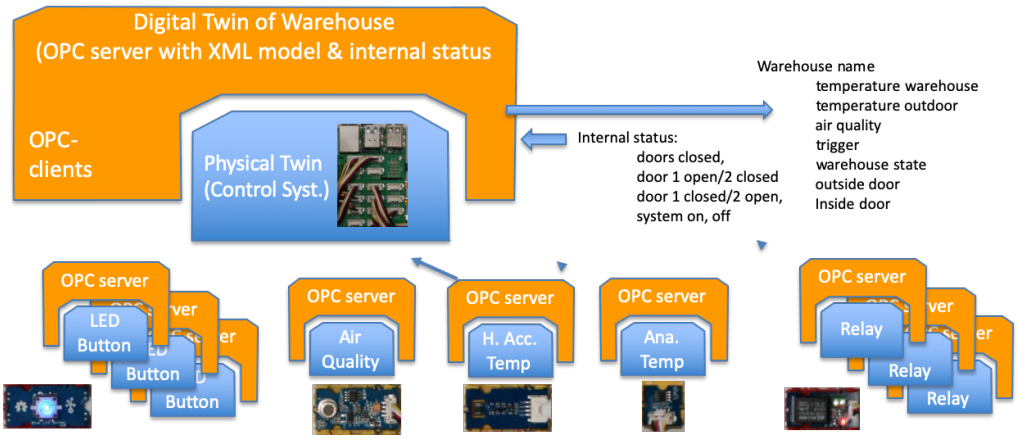


Figure 9 Nested Digital Twin with data elements.

available data without the need to see all data. Similar as you driving a car without knowing and having the access to all the internals, the manufacturing secrets as costs of components.

But whereas a product has an owner, a manufacturer, a user, etc., data can be copied over and over again. Here data ownership is more difficult to define. Only copyrighted data, i.e. where human creativity is involved as e.g. drawings, can have the copyright owned by a legal entity. And for certain data sets there exists a databank act. For private data Europe has the notion of data sovereignty in the form of the GDPR. As of today, in industry there is no legal binding concept a data ownership and/or sovereignty. There only exist contractual relations on the sharing and use of data between parties. Only the producer of the data has sovereignty control over whether to share the data with others. But e.g. sensor data has no copyrights and once shared without a contract other can do what ever they want.

5.2.2 Industrial data twins and data identifiers

Each physical object can have a digital representation in the form of a data representation as simple as just a number or as complex as a large data set with drawing, manufacturing recipes and use history. In general, it will have a name and a digital identifier too. In figure 9 a physical object or asset with its digital twin as a kind of administrative shell around it is symbolized.

An object can consists of subpart or a group of objects can form a larger object. This results in a hierarchy of digital twins each with their own data and links to other digital twins. In particular similar objects could each share the same design and their own instantiation. In that case, these digital twin instantiations (DTI's) share a common design (see figure 10 with the DT on the left and the DTI's on the right).

The design part of a digital twin could be a set of drawings, software, recipes etc. This design and the digital twin (of the design) (DT) is owned by a legal entity. And the physical objects can be owned by (other) legal entities. But what about ownership of the digital twin instantiation of each object with e.g. the historical (sensor and use) data?

5.2.3 Confidential, shareable, shared and public data sets

Each digital twin of an object, a system or even an organization can have confidential information that is not (to be) shared by anyone else. The owner of an object can also decide that certain data can be shared with other legal entities with whom the owner decides to

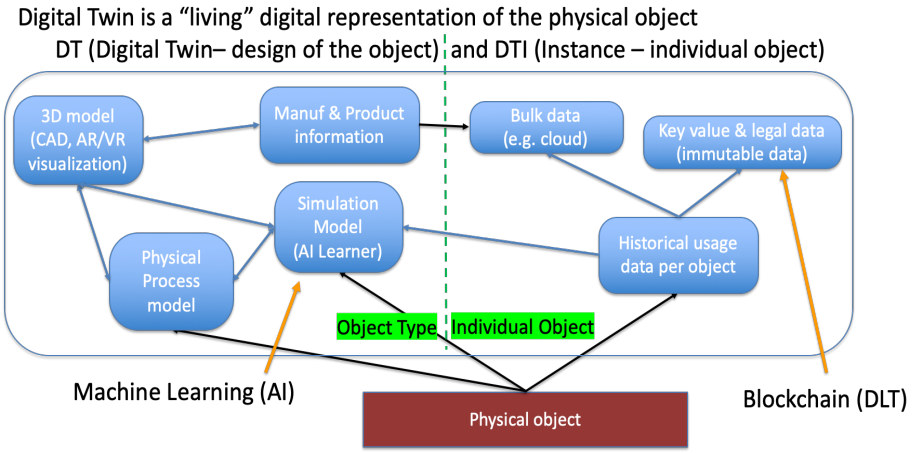


Figure 10 Digital Twin with its DT and DTI (instance).

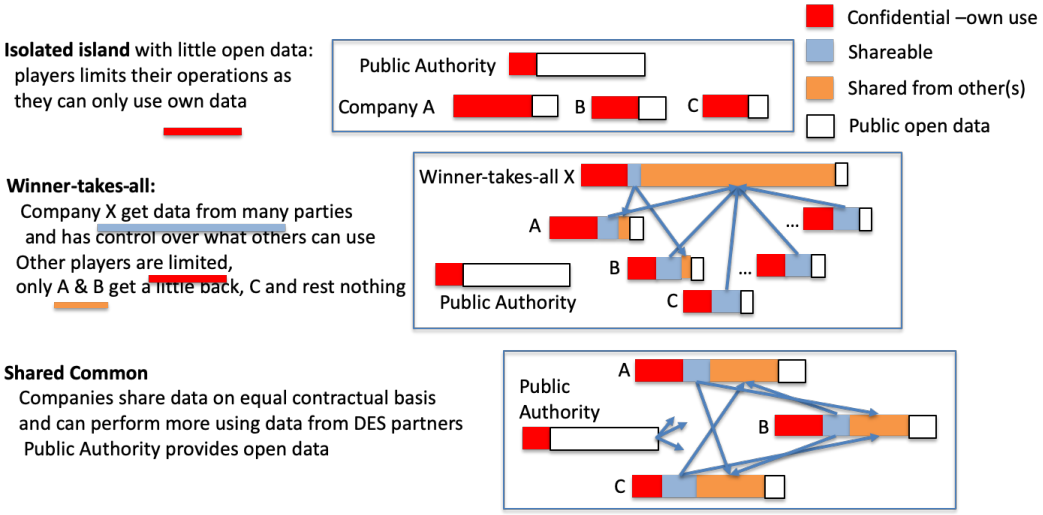


Figure 11 Anatomy of a Data EcoSystem.

share data with. This requires both a technological standard to exchange the data as well as a legal contract to specify the conditions of sharing. One can also receive data made shareable by one or more other parties, in a similar way the owners made some data shareable to others, again using tech standards and legal contracts). Finally, there might be public data that can be used too, see also figure 104 on the anatomy of three different kinds of data ecosystems illustrated as isolated islands, winner-takes-all and shared common. The last one is the preferred one in case one wants to keep a certain data sovereignty.

The result is that around every physical object, systems and even organizations a digital twin can exist with data and that these digital twins can form a data ecosystem. In industrial cases that data ecosystem will be related to physical objects, digital identifiers, data sets and relations between data sets where the digital twins can exist and be copied unlimited times in all kind of subsets while there is only one instant of the object.

Different players in a data ecosystem can have different views of the data and therefore different digital twins of the same object. Of course, by expanding the own view of the digital twin data with data shared by others a better and often more valuable digital twin can be realized.

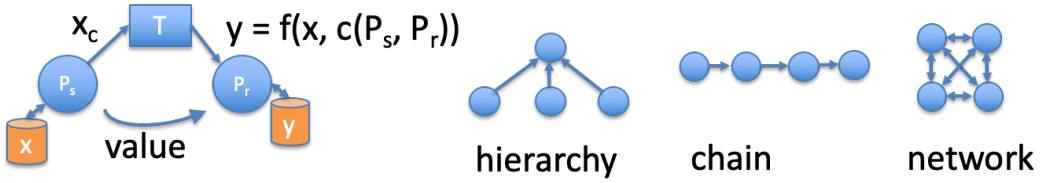


Figure 12 Different Data Ecosystem architectures.

To function in a data ecosystem, it is therefore sensible to make some or all of the own data on that DT object and/or DT instances shareable to others too. Depending on the business parties can agree to keep as much confidential (not shared) and minimize shareable data. In an extreme case one party could try to collect as much data from the others and maneuver itself in a data monopoly position. It depends on the parties how much they are willing to share, the required contracts and the technical interfaces they select and use for data exchange.

5.2.4 A Data Ecosystem function

To be able to share shareable data between parties in an ecosystem one needs, next to technical data-communication standards and legal contracts a software function. With X the data and C the legal constraint put on the data exchange, we call the entity that controls the exchange of data T (transmit). And with Ps (or S) the sending party and Pr (or R) the receiving party, T performs a function f(X, C(S, R)). T receives data X from an Ps under specified constraint instruction. See figure 12.

T can be distributed as in e.g. IDS over all parties or T can be a clearing house with processes the data as specified. Any party P could have a sending and a receiving side. In a data ecosystem there can be many S's and R's. An unbalanced situation is when there are many S and only one R, in particular if T is completely under control of the legal entity R. A tightly coupled chain might be when a flow existing from S to R follows by R to O, O to P, etc. In an hierarchical system many parties send and receive data up- and downwards and not so much side wards. Finally in an fair data-ecosystem all parties can exchange data with everybody else.

5.2.5 The GSM model

The GSM association runs a data ecosystem since the 1990-ties. With mobile telephony 1 users could not roam. With GSM a user can roam to a geographical different network of another GSM operator as in cross border travel. Because of the roaming agreement a GSM operator can provide more services then in generation 1 mobile telephony networks. But GSM operators need to share data on the roaming users. This is done by the concept of home and visitor location centers. In essence this is data-ecosystem that provides its participants more services then without it. See figure 106

In the home location center the user, the GSM nr (and device ID (IMEI)), the used service and the billing information is known. In the visitor location center only the device ID with the usage of mobile voice time and data traffic is known. This information is shared with the home GSM who then pays a bulk fee to the visited GSM operator and adds up the received usage data to the customer usage to finally produce a bill. Next to the realtime

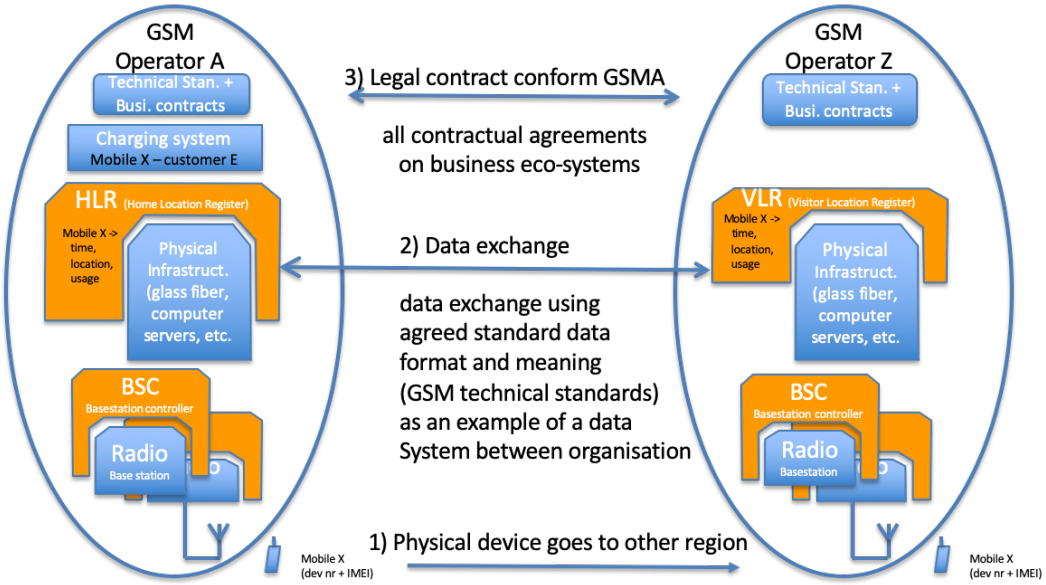


Figure 13 GSM network.

exchange of visiting or roaming data of the device there is the contractually data exchange on the total usage of visiting devices. Notice that te home GSM operators has personal data as customer name, billing info and GSM nr. The visiting GSM operator only has the device ID. When ever a device roams it device ID indicate the home GSM operator and visiting operator informs the home location register in which visitor location the device currently is. When the mobile phone number is called (in the home location visitor), the GSM home operator routes the call to the visiting GSM operator.

In the GSM case the identifiers are the home/visitor location centres, the GSM device ID each linked data sets as customers, device usage, etc. The transport T has three simple functions: updating the device visitor location (in realtime) to the home, forwarding calls to the device in the visiting network and sending the usage at aggregated level to the me too.

5.2.6 Discrete manufacturing supply chain

This example is a first-tier supplier network where orders from the OEM-er are sent to a supplier including o.a. drawing information and where finished subparts are sent to the OEM to built the product. Without the order and, if it is not a standard component, the drawing information, the supplier cannot produce. Often parties try to keep as much data as possible confidential, but more and more data is sent to the OEM-er, often on demand of the OEM-er for traceability, quality monitoring and control of the (realtime) logistics in the chain. Sometimes an independent player T could be used to assemble business information as market share and relative quality information enabling suppliers to upgrade their performance. In other cases the OEM can give feed back information to the supplier performance.

As with the GSM example, here to digital identifiers, but also digital twins and a three level physical object, digital data exchange and legal contracts are encountered.

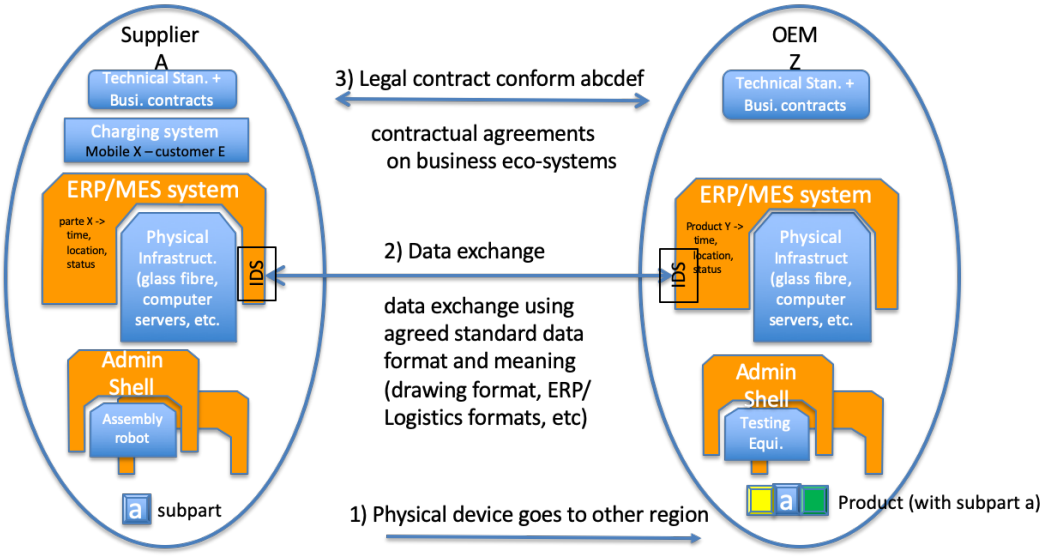


Figure 14 Supplier - Original Equipment Manufacturer network.

5.2.7 Food traceability

More and more industrial chains demand extensive traceability. Examples are the automotive and aircraft industries, but also the food industry. In figure 1 a warehouse with sensors was shown. In that case it is a part of chain where food is stored under certain temperature and humidity conditions. The warehouse operator uses the data to optimize the energy consumption, but the same data can also be shared or "sold" to the food owner to monitor the storage conditions for traceability and quality control. But sharing the data, with or without payments, the whole (data)eco system can perform better.

5.2.8 Predictive Maintenance

Predictive Maintenance is similar to the food traceability use case. In this case a maintenance service party monitors the condition of equiment to predict a stop for maintenance. Often the operator of the system uses the data for (real-time) control at the same time. By sharing the data with another party that party can combine more data for others to improve its prediction algorithms. The specifics of this use case is that these environment tend to require strict security and access control.

Figure 15 shows the case where OT (operational technology) has shielded data access in its own subnet and data cannot be sent directly to others. In this case the data is sent by OPC-UA (IEC standard) protocol through a firewall to a server to be stored there and made suitable to be further processes outside the operational production environment. These two level data ecosystems often restrict other legal parties to have an active proces directly communicating through the customers network to their own remote services support networks. OPC-UA is used for equipment communication within an organization. Other standards as the IDS (international data spaces) can be used for data exchange between different organization using the IDS technical and contractual agreements.

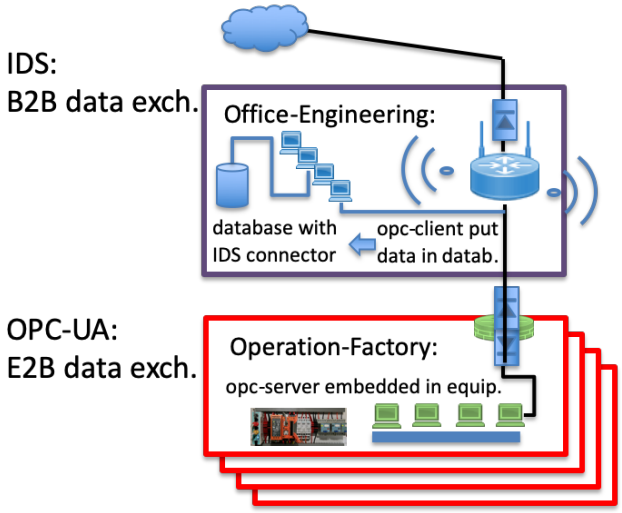


Figure 15 data ecosystem inside and outside a factory.

5.3 Use Cases from the Smart Cities Domain

Cinzia Cappiello (Polytechnic University of Milan), Bernadette Farias Lóscio (Federal University of Pernambuco), Avigdor Gal (Technion – Haifa, IL), Fritz Henglein (Univ. of Copenhagen, DK & Deon Digital – Zürich, CH)

The Smart city concept has various definitions that are associated with viewpoints that range from the people perspective to a more technological perspective. Such latter point of view suggests that smart cities mainly focus on adopting the next-generation information technology to "all walks of life, embedding sensors and equipment to hospitals, power grids, railways, bridges, tunnels, roads, buildings, water systems, dams, oil and gas pipelines and other objects in every corner of the world, and forming the "Internet of Things" via the Internet"[103]. Smart Cities data collected by heterogeneous devices and data sources are strategic to perform analysis, to extract relevant patterns, to detect inefficiencies, and to propose innovative solutions to improve the quality of life of citizens.

5.3.1 Scenario description

Goal: Definition of an ecosystem that includes several cities (learning from one city can help to accelerate the process in another city).

The Smart Cities DE, presented in Figure 16, can be seen as a composition of Data Ecosystems. Each city has its own DE and they can exchange data and knowledge. Each DE is composed by Data Providers and Services. The Data Ecosystem of City 1, for example, is composed by five Data Providers, which can also play the role of a Data Consumer. Data from different providers can be used by a Data Consumer to offer more advanced service. In this scenario of data mobility, Data Provider 1 and Data Provider 3 provide bus data to Data Provider 2, which will perform time predictions based on these data. In a similar way, Data Provider 4 receives bus data from Data Provider 1 as well as time predictions from Data Provider 2. Data Provider 4 uses these data to develop a better location planning

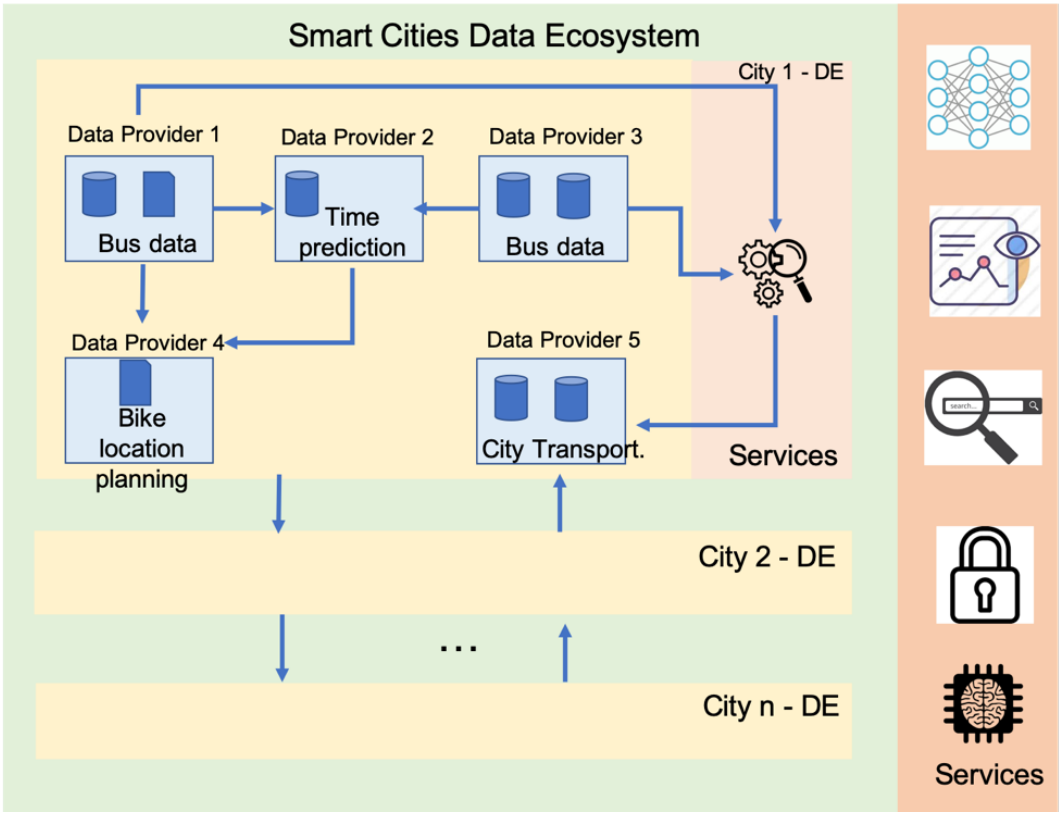


Figure 16 Smart cities data ecosystems.

for bikes to rent. Finally, Data Provider 5 uses the services offered by the DE to provide information to citizens.

In such ecosystem, we have to clarify that a city can learn from another city but it is difficult that a city might inherit or use the same applications/services of another city. In fact, most of the smart city applications are context- aware and tailored for the city needs.

5.3.2 Challenges

- General challenges
  - It is important to guarantee that people use data in the correct way: data usage constraints have to be defined.
  - Looking at the data infrastructure, it is difficult to understand which are the boundaries of a node. It might be adopted a data provider-oriented approach in which a node is a data provider within its own data sources.
- Challenges related to derived data issues
  - In the data ecosystem of a city, in general a data provider can be the owner of raw data sources or can provide derived data created transforming data gathered by other data providers. In summary we can distinguish between raw data and derived data. It is important to find a way to describe and manage derived data.
  - How do we define the usage constraints of the derived data? It might happen that raw data are regulated by usage constraints but derived data do not contain personal data and vice-versa. Besides, integration between two sources could reveal personal data.

  - The code related to the applications/services should be represented. Also the code should be protected
  - View maintenance: model improvement or data improvement.

- Challenges related to Quality assessment
  - In this architecture, exchanged data should be high quality data in order to exploit the potential data value. It is necessary to define who should be in charge of the quality assessment. In fact, data quality assessment could be centralized and in charge of the platform through a platform service also using crowdsourcing.

6 Manifesto

The Dagstuhl Manifesto is an agreed result of the Dagstuhl Seminar "Data Ecosystems: Sovereign Data Exchange among Organizations" that took place on 23-27 September 2019 in the Schloss Dagstuhl Leibnitz Zentrum fur Informatik, that joined forces of the researchers, practitioners and experts from Europe and worldwide, with the goal to discuss new challenges in building data ecosystems supporting sovereign data exchange among organisations involved into whole data value chain. We understand that:

1. Data exchange among organizations is a key enabler for the digital economy of the future.
2. A secure, reliable and performant data exchange infrastructure is a basis for operating sustainable data ecosystems supporting the whole data value network involving data providers, data consumers, and service providers.
3. Enabling assessment and awareness of data quality are core requirements in a data ecosystem.
4. Organizations and individuals must be regarded as data sovereigns entering into data ecosystems according to agreed contracts.
5. In a data ecosystem, data should be considered both an economic asset and a tradable commodity according to specified conditions of use.
6. Defining metadata is necessary for enabling a variety of operations on data and with data, along the whole data value network.
7. Defining semantic interrelationships among data sets is a key problem/task in creating scalable data ecosystems supporting effective data exchange.
8. Defining a generic data ecosystem architecture is necessary for supporting interoperability at multiple ecosystem levels, including data sources, semantics, applications, workflows and processes, governance and economics.
9. Economically viable data ecosystems architectures should leverage successful experiences from similar distributed systems such as GSM employing context roaming between independent domains.
10. Future data ecosystems should serve the public interests, in particular by supporting data related projects to address the UNDP 17 Sustainable Development Goals (SDG) .

7 Statements of the participants

7.1 Cinzia Cappiello (Polytechnic University of Milan, IT)

My research focuses on Data and Information Quality. In particular, in the last years, I addressed issues related to the Data and Information Quality assessment and improvement in the context of service-based applications, Web applications, Big Data and IoT. Data Quality is defined as fitness for (intended) use, that can be seen as the capability of a data set to satisfy users' requirements [6]. This definition suggests that quality is subjective: a data set that is appropriate for an application/user might not be suitable for another one. For this reason, currently, I am working on the design of an application-aware data quality assessment platform: data quality should be evaluated by considering the actual usage of the considered data set. Moreover, in this field, most of the literature contributions propose approaches for structured data sets. In data ecosystem, the increasing volume and variety of the data sources needs the definition of new data quality methods. Such methods should be designed by considering the type of source (e.g., data streams vs traditional db) and the type of data (e.g., text vs numerical). In fact, these two variables impact on the dimensions to evaluate and on the metrics to use. In summary, in order to enable an automatic evaluation of the quality level of the data sources it is necessary to define an adaptive data quality assessment service able to trigger the appropriate mechanisms by considering the characteristics of the sources and the application/user that requested data.

7.2 Ugo de'Liguoro (University of Turin, IT)

My interests are to logic and computation. Central are both pure and typed λ-calculus. I have also investigated other formalisms, like the ζ-calculus and the π-calculus, considered as theoretical calculi modeling object-oriented and concurrent programming languages respectively. Key methodologies in my work are type systems and denotational semantics, that nicely correspond each other especially in the case of intersection type assignment systems. The latter express functional properties of λ-terms seen as programs, and I have worked to include in the theory also non functional aspects, like non determinism and control operators. I believe that intersection types have a great potential to model both behaviour and data; an experience in that direction is the participation to a project of program synthesis from components in Dortmund, based on intersection types and combinatory logic. In a related field I have been working on session types and contract theory. This is about modeling protocol compliance of multiple principals interacting through a network, enjoying safety and liveness properties. A particular concern in this work is checking for consistency of local protocols to form a well behaved system, and the adaptation of each to the others when considered as components independently built and specified.

7.3 Yuri Demchenko (University of Amsterdam, NL)

The fact that data has a value is commonly recognised. However, data value is different from those associated with the consumable goods. There are a number of initiatives to create data markets and data exchange services. Well established business models of paid or commercial data(sets) services such as data archives are based on the service subscription fee. Quality of datasets in many cases is often assessed by independent certification body or

based on peer review by expert. However this model does not provide a basis for making data an economic goods and enable data commoditisation. Another important development in making the best use of research data is based on wide implementation of the FAIR (Findable – Accessible – Interoperable – Reusable) data principles, which are widely supported by research and industry. However, emerging data driven technologies and economy facilitate interest to making data a new economic value (data commoditisation) and consequently identification of the data properties as economic goods. The following properties proposed in [32] leverage the FAIR principles and are defined as STREAM for industrial and commoditised data: [S] Sovereign - [T] Trusted - [R] Reusable - [E] Exchangeable - [A] Actionable - [M] Measurable Other properties to be considered and necessary for defining workable business and operational models: nonrival nature of data, data ownership, data quality, measurable use of data, privacy, integrity, and provenance. Defintion of the data properties as economic goods must be supported by creating consistent and workable models for data exchange and commoditisation, to facilitate creation of new value added data driven services. Consistent data pricing and data markets models are equally important for government funded and sponsored research, open data and governmental data. The proposed Open Data Market (ODM) model is based on the adoption of the IDSA Architecture and data sovereignty principle [87]. The ODM must be based on relevant industry standards and provide secure and trusted data exchange between data market actors: data producers/owners and data consum-ers, services and applications developers and operators. A functional data market model and architecture should include multiple components such as the secure trusted data market infrastructure as well as regulatory basis. The proposed open data market model would be decentralized and allow creating virtual private market instances to support data exchange between peers or group of peers. This gives the advantage that network nodes, data sellers and data buyers, receive the full benefits of the data market while retaining full control of their data. This is an important requirement for data markets for industrial data where companies, the data owners, want to retain control over their data, maintaining data sovereignty. The ODM research evaluates the blockchain technology to enable an open controllable trusted data market environment for secure and trusted data exchange, and support data value chain (provenance) and create a bias for data monetisation [13].

7.4 Elena Demidova (Leibniz Universität Hannover, DE)

Recently, we have been involved in several projects focused on analysing data for urban mobility use cases. Examples include prediction of impact of planned special events on traffic [106], identification of structural dependencies in urban road networks [107], or analysis of driver behaviour. Realisation of such use cases requires integrated analytics of data originating from different domains (map data, traffic information, car trajectories, car sensor data, city infrastructure data, population statistics, people movement, etc.), as well as different sources and owners. Currently, most of the data relevant for such analytics is locked by different organisations (e.g. companies or city municipalities), making it hard to obtain the overall picture and build accurate prediction models. In this seminar, I would like to share

[13] https://datapace.io/datapace_whitepaper.pdf

experiences we have collected in several related projects (including Data4UrbanMobility[14] [108], Simple-ML[15][53] and CampaNeo[16], leading to challenges related to collecting such data, and making use of the data in the analytical use cases more transparent to data owners.

7.5 Boris Düdder (University of Copenhagen, DK)

Boris Düdder's core interests are in formal methods of software engineering and distributed systems, for example, by employing logical methods, rewriting systems, and model checking. His interests concerning the Dagstuhl Seminar are methods for automatic configuration of exchange of sovereign production data of manufacturers and suppliers securely while guaranteeing ownership, confidentiality, and privacy. He can contribute his expertise and industrial experience in formal methods for synthesis, analysis, and verification of software for industrial production systems and supply chain (including logistics) as well as distributed ledger technology, e.g., for proof of provenance applications and finance. Industry experience includes logistics, whiteware and aircraft manufacturers.

7.6 Bernadette Farias Lóscio (Federal University of Pernambuco, BR)

In the last years, I have been involved in several projects related to Open Data and Data on the Web. One of my main interests concerns how to share data on the Web in a proper and sustainable way. The growing interest in sharing data on the Web gives raise to several challenges related to important subjects including data provenance, data privacy and data access. In order to help data providers and data consumers to face those challenges, the W3C Data on the Web Best Practices Working Group proposed a set of 35 Best Practices, which cover different aspects related to data publishing and consumption, like data formats, data access, data identifiers and metadata. I was part of the DWBP working group and I am one of the editors of the Data on the Web Best Practices recommendation. This experience was very rich and I could have the opportunity to identify several open research challenges related to Data on the Web. One of these challenges concerns the creation of sustainable Data Ecosystems.

While Data Ecosystems are gaining importance, several ecosystems are not sustainable and consequently the effort spent by their actors end up not being properly used or forgotten. The lack of communication and cooperation between data producers and consumers is one of the main obstacles moving towards sustainable Data Ecosystems. Moreover, designing, developing and further maintaining systems for Data Ecosystems are not trivial. Recently, I have advised several students in subjects related to Data Ecosystems, including a metamodel to represent data ecosystems, a metadata curation framework for data ecosystems and a framework to assess data ecosystems health. In this seminar, I would be glad to share my experiences on these subjects and I am willing to learn from the experiences of other participants.

[14] http://data4urbanmobility.l3s.uni-hannover.de
[15] https://simple-ml.de/
[16] https://www.l3s.de/de/projects/campaneo

7.7 Avigdor Gal (Technion – Haifa, IL)

My research focuses on effective methods of integrating data from multiple and diverse sources, in the presence of uncertainty. My current work zeroes in on schema matching – the task of providing communication between databases, entity resolution – the task of identifying data elements that relate to the same real-work entity, and process matching – the task of aligning process activities. In the context of this seminar, I have contributed to the design of an intelligent data lake, one in which integration is seamlessly generated and tested with the assistance of experts, heuristics and machine learning algorithms

7.8 Sandra Geisler (Fraunhofer FIT – Sankt Augustin, DE)

My field of expertise and research comprises several aspects of data management. Especially, my work is concerned with techniques in the area of data stream management, big data in general, data quality, data lakes, meta-data management, data integration, and semantic web technologies. Furthermore, I have a strong background in the medical domain and medical informatics.

Currently, I am working on projects in the context of the Fraunhofer Medical Data Space [17]. The Medical Data Space can be viewed as a distributed data ecosystem which has specific requirements stemming from the challenges of working with health data. It is especially concerned with data sovereignty, data security, and data privacy striving for compliance to data protection laws on various levels. The data in the Medical Data Space remains in custody of the data owner and is only exchanged in a controlled and secure manner. For example, data important for the follow-up care of patients after a surgery may be shared between several parties involved in the care process, such as the general practitioner, the outpatient nursing service, the hospital, and of course the patient herself. But, every party may only need a specific part of the data, or should not be allowed to access all of the data, or should only see results produced by algorithms using the data as input. Furthermore, data quality plays a crucial role as the data may be used to support processes which may have indirect or direct impact on the health of a person. Also data sources with potentially erroneous data, such as health monitoring devices, are often involved. But, controlled and automated data collection of the users may also help to elevate DQ which in consequence will also lead to a higher data value and quantity for health care services or other data consumers, such as research studies. Finally, challenges regarding the integration and interpretation of the distributed and very heterogeneous data have to be tackled to provide a use case specific and user-friendly view on the data.

Based on my research interests and current project work, I can contribute to the seminar in terms of the discussion of applications from the health care sector and their specific challenges and requirements focusing on data sovereignty, data quality, and data integration.

[17] https://www.medical-data-space.fraunhofer.de

7.9 Benjamin Heitmann (Fraunhofer FIT – Aachen, DE & RWTH Aachen, DE)

The data market places of the future will be enabled by technologies which protect the value of data for all participants of a market place. In my experience, too often the real value of data or other digital assets for a stakeholder is neglected, as processes and algorithms can currently be better protected. In order to protect the value of data in a market place, the three goals of security, privacy and data sovereignty need to be fulfilled.

Security means protecting the digital assets which an organisation controls and which might be valuable to an attacker. Examples of assets which usually are secured, include domain-specific knowledge collections such as life-science data sets, financial records, records of human resource data, product sales data, and instance data for business processes. Failure to protect the digital assets of an organisation results in losing control of knowledge, processes and customers, as competing organisations could gain access to the same assets.

Privacy means protecting the data of individual users, which is required as input for many digital value chains. Examples of private assets include digital health records, purchase histories for e-commerce sites, historical data about consumption habits for news, movies and music. Failure to protect the personal data of individual users will result in unwillingness or reduced incentives for users to share or sell their data. This will in turn disable any processes which require input data from users to generate more value for an organisation.

Data sovereignty means protecting the digital assets which an organisation or individual controls and which have been sold or shared willingly with another entity. Examples of assets which might be shared or sold without losing control include R&D data sets for manufacturing, R&D data sets for discovering new gene interactions, personal data about historic fitness activities or health indicators such as blood pressure. Failure to protect assets after sharing or selling them will result in loss of control over the purpose of using the data, and in decreased incentives to share or sell the data with additional entities in the future.

In future data market places, the data protection needs to fulfil the requirements of all stake holders at the same time: data generator, seller and buyer. In addition, usually all three goals of security, privacy and sovereignty need to be reached at the same time. This requires technologies which enable **enforceable guarantees for protecting data during processing**. In my experience, some of the technologies with the highest potential come from new advances in information processing and encryption. The most notable new developments are in the areas of secure multi-party computation (SMPC), homomorphic encryption (HE), and differential privacy, anonymisation and synthetic data generation.

Together with new architectures, business models and legal frameworks, the listed technologies for enforceable guarantees to protect data during processing can enable sustainable data markets which fulfill the requirements of all participants.

7.10 Fritz Henglein (Univ. of Copenhagen, DK & Deon Digital – Zürich, CH)

Fritz Henglein is Professor of Programming Languages and Systems at DIKU, the Department of Computer Science at the University of Copenhagen (UCPH) and Head of Research at Deon Digital AG, a Zürich/Copenhagen-based start-up developing secure and scalable digital contract technology for both decentralized (blockchain, distributed ledger) and centralized systems. His research interests and contributions are in semantic, logical and algorithmic aspects of programming languages, functional programming, domain-specific languages,

digital contracts, reporting and analytics, smart contracts and distributed ledger technology, with applications in enterprise systems, business processes, high-performance computing, probabilistic programming and decentralized systems, including blockchain and distributed ledger systems.

Key desiderata of (next-generation) distributed ledger (DL) systems are tamper-proof, privacy-respecting recording of real-world and business events together with their evidence; secure distributed storage and transfer of assets such as money, assets, securities and (digital twins of) physical resources; automatically managed and enforced contracts that provide an effective and auditable basis for privacy preserving collaboration, analytics and planning; a balance of availability, consistency and network failure tolerance tuned to specific use case characteristics; and, most importantly, organizational decentralization to minimize the need for and competitive advantage of a dominant and controlling platform provider. Such a DL system facilitates tracking of digital resources, including valuable data, across organizations as well as transfer (and revocation) of control over them. Ongoing research, development and commercial application (at Deon Digital) of smart digital contract technology on existing commercial distributed ledger systems indicates that this may facilitate effective cross-organizational data exchange by facilitating unforgeable auditable proofs of authorized data use.

7.11 Matthias Jarke (RWTH Aachen University and Fraunhofer FIT, DE)

Data-driven machine learning methods are typically most successful when they can rely on very large and in some sense homogeneous training sets in areas where little prior scientific knowledge exists. Production engineering, management, and usage satisfy few of these criteria and therefore do not show many success stories, beyond narrowly defined specific issues in specific contexts. In contrast, the last years have seen impressive successes in model-driven materials and production engineering methods, these methods lack context and real-time adaptivity.

Our vision of an Internet of Production, pursued in an interdisciplinary DFG-funded Excellence Cluster at RWTH Aachen University, addresses these shortcomings: Through sophisticated heterogeneous data integration and controlled data sharing approaches, it broadens the experience base of cross-organizational product and process data. At the method level, it interleaves fast "reduced models" from different engineering fields, with enhanced explainable machine learning techniques and model-driven re-engineering during operations.

As a common conceptual modeling abstraction, we investigate Digital Shadows, a strongly empowered variant of the well-known view concept from data management. The idea of Digital Shadows dates back to Platon's famour Cave Allegory but which he illustrated the limitations of all human knowledge – we can always only see partial perspectives on the world. Fifty years of data management research confirm that the growth of data has always outpaced our ability to deal with them, and we expect it to stay this way. Therefore, we see strong limitations for the currently fashionable Digital Twins when real-time and large scope are relevant, and propose to circumscribe them by well-structured collections of Digital Shadows. Besides enabling real-time monitoring and control at an abstract level, Digital Shadows are also well-suited for bridging interdisciplinary boundaries and communication problems between research and practice. Several initial experiments indicate the power of this approach but also highlight many further research challenges.

7.12 Jan Jürjens (Universität Koblenz-Landau, DE)

AI uses scientific methods, processes, algorithms, and systems to gain knowledge and insights into data that exists in a variety of formats. Characteristic of the area is on the one hand the high significance that it has. Based on the results obtained, numerous important decisions are made concerning the individual or society as a whole: diagnoses, therapies, credit decisions, spatial planning, etc. On the other hand, AI is characterized by the iterative and empirical-heuristic approach by which knowledge is extracted and decisions are derived. From the point of view of the provider of the data on which the AI-based analysis is performed, it is important that the Data Sovereignty is preserved in the context of the data analysis, which is closely related to the following aspects:

- Data security
- Data privacy
- Transparency and explainability
- Fairness / non-discrimination

Unfortunately, it is a significant challenge to be able to demonstrate whether or not these aspects of Data Sovereignty are satisfied in a given situation involving AI based analysis. In fact, it is already a challenge to just describe "correct behavior" of an AI based system, because its results are usually not predetermined and can only be obtained through the data analysis process, so it is in general not clear what "correctness" means in this context. The challenge is also that the behaviour of a self-learning algorithm depends on the training data to which it is continuously being subjected and therefore cannot be determined by only considering the algorithm itself. We thus need a verification approach which can be efficiently parameterized over the possible effects of the self-learning process without having to re-do the complete verification whenever the self-learning leads to a change in behaviour. The proposed talk discusses these challenges and presents an approach that supports the analysis of Data Sovereignty aspects in the context of AI-based systems based on the tool-based analysis of software design models in UML, which supports change-based verification and can thus deal with the different variants of algorithm behaviour arising from self-learning. The talk is based on work done as part of University of Koblenz' research priority programme "Engineering Trustworthy Data-intensive Systems" as well as within the context of the "International Data Spaces" initiative at Fraunhofer ISST.

7.13 Maurizio Lenzerini (Sapienza University of Rome, IT)

Data interoperability refers to the issue of accessing and processing data from multiple sources in order to create more holistic and contextual information for improving data analysis, for better decision-making, and for accountability purposes. In the era towards a data-driven society, the notion of data interoperability is of paramount importance. Looking at the reseach work in the last decades, several types of data interoperability scenarios emerged, including the following.

1. In Data Integration, we have a multitude of information sources, and we want to access them by means of a global schema, that somehow accomodates an integrated view of all data at the sources [35, 75].
2. In Data Exchange, we have a source databas, and a target database, and we want to move the data from the source to the target according to some specified criteria [5, 66].

3. In P2P Data Coordination, we have a network of information nodes (peers), and we want to let them communicate to each other in order to exchange data or queries [20, 76].
4. In Ontology-Based Data Management (OBDM), we have a collection of data sources and an ontol- ogy representing a semantic model of the domain of interest, and we want to govern (i.e., query, update, monitoring, etc.) the data at the sources through the ontology, rather than by interacting directly with the sources [29, 76].

A fundamental component of all the above data interoperability frameworks is the mapping. Indeed, put in an abstract way, all the above scenarios are characterized by an architecture constituted by various autonomous nodes (called databases, data sources, peers, etc.) which hold information, and which are linked to other nodes by means of mappings. A mapping is a statement specifying that some relationship exists between pieces of information held by one node and pieces of information held by another node. Specifically, in Data Integration the mappings relate the data sources to the global schema, in Data Exchange they relate the source database to the target database, in P2P Coordination they relate the various peers in the network, and in OBDM they relate the various data sources to the ontology. In the last years, many papers investigate the notion of mapping, from various points of view, and with different goals (see [67] and references therein). By looking at these papers, one could argue that one of the most important role of mapping is to allow reformulating queries expressed over a node into queries expressed over other mapped nodes. Such reformulation task is crucial, for example, for answering queries expressed over the global schema in a data integration system. Indeed, to compute the answer, the system has to figure out which queries to ask to the data sources (where the real data are located), and this is done by a step that we call direct rewriting: rewrite the query over the global schema in terms of a query over the data sources. A similar task has been studied in the other data interoperability scenarios. In OBDM, for instance, given a user queries expressed over the ontology, the aim is to find a direct rewriting of the query, i.e., a query over the source schema, that, once executed over the data, provides the user query answers that are logically implied by the ontology and the mapping. While the notion of direct rewriting has been the subject of many investigations in data interoperability in the last decades, in this paper we aim at discuss also a new notion of rewriting, that we call inverse rewriting. The importance of this new notion emerges when we consider the following task in the OBDM scenario: Given a query q over the sources, find the query over the ontology that characterizes q at best (independently from the current source database). Note that the problem is reversed with respect to the one where the traditional (direct) rewriting is used: here, we start with a source query, and we aim at deriving a corresponding query over the ontology. Thus, we are dealing with a sort of reverse engineering problem, which is novel in the investigation of data interoperabilty. We argue that this problem is relevant in a plethora of application scenarios. For the sake of brevity, we mention only three of them. (1) Following the ideas in [25], the notion of reverse rewriting can be used to provide the semantics of open data and open APIs published by organizations, which is a crucial aspect for unchaining all the potentials of open data. (2) Although the architecture of many modern Information Systems is based on data services, that are abstractions of computation done on data sources, it is often the case that the semantics of such computations is not well specified or documented. Can we automatically produce a semantic characterization of a data service, having an OBDA specification available? The idea is to exploit a new reasoning task over the OBDA specification, that works as follows: we express the data service in terms of a query over the sources, and we use the notion of reverse rewriting for deriving the query over the ontology that best describes the data service, given the ontology and the mapping.

(3) It can be shown that the concept of reverse rewriting is also useful for a semantic-based approach to source profiling [2], in particular for describing the structure and the content of a data source in terms of the business vocabulary.

7.14 Wolfgang Maaß (Universität des Saarlandes – Saarbrücken, DE)

Intelligent services using Artificial Intelligence methods are tools for creating, modifying and merging data products from different industries, in particular industrial production systems (Industrie 4.0). We take an extended view on data products, which includes (a) data as statements about a domain, (b) software, and (c) models including conceptual models, machine learning models, and semantic models. In our research we develop AI algorithms and architectures for distributed (IOT) environments. These technologies are used to build domain-specific smart service systems. For example, Smart Dialog Services are used to make data products and derive decision recommendations based on machine learning mechanisms accessible to human experts (explainable AI and responsible AI). A further focus is the investigation of means for the automatic evaluation of data products as input for the financial reporting of digital assets. Our research has been conducted in various fields, such as industrial manufacturing, health care/medicine and media. While the media industry and medicine have been struggling with the challenges of digital transformation for about 30 years, industrial production is at the beginning of understanding digital products as an independent economic asset class.

7.15 Paolo Missier (Newcastle University, GB)

Data marketplaces are becoming ubiquitous, but for the most part, they assume (i) static data, and (ii) a trusted environment where data trading takes place. Both these assumptions introduce limitations in the potential of individuals to exchange their own personal data with other parties. Data streams that originate from Internet of Things (IoT) devices, often placed in people's premises (house, vehicle, or own body using wearables), are increasingly viewed as tradeable assets with value not only to the device owners, but also with resell value, i.e., to third party buyers. We envision a marketplace for IoT data streams that can unlock such potential value in a scalable way, by enabling any pairs of data providers and consumers to engage in data exchange transactions without any prior assumption of mutual trust. We have recently proposed one such next-generation marketplace model, where the requirement for trust in data exchange is fulfilled using blockchain technology and specifically Smart Contracts. I am interested in discussing such marketplace model within the context of this Dagstuhl Seminar, as I believe it has the potential to become a cornerstone of data ecosystem where organisations exchange their data reliably and in real-time, subject to legally binding trading agreements, and while providing incentives for fair trading.

7.16 Boris Otto (Fraunhofer ISST – Dortmund, DE & TU Dortmund, DE)

The proliferation of digital technologies such as cloud platforms and cyber-physical systems in the manufacturing industry enables process and service innovation in production and supply networks. Sharing data among network partners in an important prerequisite for

leveraing this innovation potential. In this context, the notion of "digital twins" has received significant attention, both in the scientific and the practioners' community. In general, a digital twin represents real world objects (such as production equipment, components, material etc.) along their lifecycle. It consists of descriptive data (e.g. the eCl@ss number of a tool machine component) as well as event data (such as temperature, vibration data etc. captured in the production process). Furthermore, the digital twin comprises both metadata and data items. A prominent digital twin information model, namely the so-called Asset Administraive Shell proposed by the German "Plattform Industrie 4.0" distinguihes between "types" and "instances". Recent digital twin scenarios are based sharing digital twin data within a production and supply network or even within an industrial ecosystem. Hence, the concept of "shared digital twins" of components, machines etc. emerges. While the fundamental concepts of digital twins in general (such as metadata, data models, distributed data storage are mature research topics, sharing digital twin data among different partners has not sufficentily be conceptualized. Moreover, taking an Information System Research perspective, a set of interesting research questions can be identified. Examples are:

- What are appropriate query approaches to retrieve shared digital twin data?
- How to distribute storing and processing of shared digital twin data between edge, fog, and cloud level? What are appropriate synchronization approaches?
- How do ensure data sovereignty of different contributers to the shared digital twin in a complex production and supply network?
- How to design functional data governance models for shared digital twin?
- How to enable access and usage control for shared digital twin data?
- How to design data provenance architectures for shared digital twins.

Based on a solid conceptulization of shared digital twins, a research agenda is required to identify research demand and outline promising research trajectories.

7.17 Elda Paja (IT University of Copenhagen, DK)

My research focuses on techniques to support the elicitation or design of socio-technical requirements (looking not only at software, but also humans and organizations, their interactions and message exchanges), requirements that come from different stakeholders, and might be conflicting with one another or impact business policies or system functionality. In my work, I have been exploring the design of socio-technical systems from different angles: security, privacy, risk, and decision-making support. The common ground of my approach is the use of conceptual modelling techniques, and in particular the use of modelling primitives to describe the rationale behind the behaviour of the various participants of a socio-technical system, and the definition and exploitation of automated reasoning techniques to support the work of requirements engineers. Contributions to the seminar

- Discussing on analogies and potential of adapting existing methods for requirements engineering for data ecosystems, treating data as first class citizens.
- Presenting recent work on consent privacy requirements for sociotechnical systems, following a multi-level approach, starting from social and organizational requirements, to business process level verification of deviations or breaches.
- Presenting recent work on consent verification monitoring, namely a formal framework to support companies and users in their understanding of policies evolution under consent regime that supports both retroactive and non-retroactive consent and consent revocation, all in a context where personal data provides important business value, e.g. in the personalization of services.

7.18 Barbara Pernici (Polytechnic University of Milan, IT)

Sharing of scientific and scholarly data is enabled by open or shared repositories in many different scientific domains. Data sharing and open data are not final goals in themselves, however, and the real benefit is in data reuse by different actors.

Focusing on reuse, the design of integrated frameworks that make it simple and effective both the upload and the retrieval of large amounts of scientific data extracted from the literature or produced by research labs, to support research, and in particular scientific model development, is a challenging issue. In particular the development of scientific models to reproduce and predict complex phenomena is a challenging task, which requires a rich set of data both for model development and validation. This task is becoming particularly critical as more and more approaches to automate model development using machine learning techniques emerge in different fields.

Starting from ongoing researches in the domain of chemistry engineering and emergency information systems, several research questions emerge and will be discussed:

- How to represent metadata in a flexible way, in particular when there is no agreement within the ecosystem of contributors on a common underlying model and new features emerge during the analysis.
- How to allow several organizations simultaneously analyze the data from different perspectives and integrate their results.
- How to assess the quality of the data and of the analysis/classification models, possibly automatically derived through deep learning mechanisms, distinguishing between assessment of quality of data (both training and input data) and other quality issues introduced by the analysis models.
- Another open issue concerns the intellectual property and access to results both of the original data and of data obtained through modeling/ simulation processes.
- Finally, the interaction of information systems experts, computer scientists and experts on the specific domain of interest requires new flexible approaches and design patterns in the development and of such systems.

7.19 Frank Piller (RWTH Aachen, DE)

One of our core research streams at the Institute for Technology and Innovation Management at RWTH Aachen University investigates the need of established corporations to deal with disruptive business model innovation and supporting organizational structures and cultures. The rise of platform-based business models (or business models for industrial data ecosystems) is one of the main drivers of change in this field. Hence, we currently have a number of research projects where we study the systematic development of these platform-based business ecosystems and the strategic positioning of an industrial company in these ecosystems. The largest among these projects is the Cluster of Excellence “Internet of Production (IoP)”, funded by the German Research Council (DFG), Project ID 390 621 612). The IoP resembles the vision of an open network of sensors, assets, products, and actors that continuously generate data. A core element hence is the (re-)use of data, digital shadows, and applications by other parties to facilitate faster and more efficient learning and analytics. The rise of platforms (business ecosystems) where these data are being exchanged and enhanced by dedicated “apps” is a central element of the IoP vision. To create value, ecosystems build on complementary inputs made by loosely interconnected, yet independent stakeholders. Among

these participants, dedicated mechanisms governing data access and privacy are required. Our research here takes an inter-firm (external) perspective: setting the right incentives for sharing deep production know-how and data while balancing value creation with value capture (sharing the rents) for all participants. Among others, we are interested in the following research questions:

- Modeling the tension between openness in value creation and control of value capture,
- Managing property rights (access, transfer, enforcement) at data, applications, and connected assets as a result of varying degrees of platform openness, and
- Definition of governance modes and design factors to generate adequate business models for the IoP that allow to maximize value appropriation for all involved actors.

7.20 Andreas Rausch (TU Clausthal, DE)

Data today plays a much bigger role than it used and now data is the new oil. The reason why data is generated and collected in such a massive amount is pretty simple. Data is the new oil and has become a fuel for new business models. However, most of the available data is related to very limited number of companies and organizations that form almost an oligopoly – the so-called GAFA companies [18] [19]: Google, Facebook, Apple and Amazon. For Small and Medium Enterprises (SMEs) it is much harder to extract the same value from their data compared to these large enterprises. On one hand it is very difficult for enterprises (Data consumers) to obtain proper data and on other hand for those who collect data (Data providers), the problem is- How to draw to additional profit from the data beyond its obvious purpose. Thus, a common data sharing platform is required where the data producers can obtain profit from their data and the data consumers can easily find data. To tackle this a new data marketplace ecosystem is required based on just technical aspects but also the social and economic aspects. Rather than conventional centralized solutions, our research focuses on solutions for enabling the data sharing without the stakeholders having to have full trust in the marketplace owner or provide.

The three main building blocks on this new proposed "Data marketplace ecosystem" are a community system, open business architecture platform and the relationship between community system and open business architecture platform. This data marketplace ecosystem is a decentralized, open and large software system, which is owned, controlled and used by a community system.

- A community system: A community system is a group of people who share a common interest but still form a heterogeneous system. The community system can be subdivided into different homogeneous subgroups.
  - Provide community- The provider community defines some rules and standards for the ecosystem to function safely. It also helps the ecosystem evolve based on the requirements of the user community.

[18] Z. Abrahamson, "Essential Data," 2014.

[19] Lucy Handley, "Amazon beats Apple and Google to become the world's most valuable brand," 2019. [Online]. Available: https://www.cnbc.com/2019/06/11/amazon-beats-apple-and-google-to-become-the-worlds-most-valuable-brand.html [Accessed: 18-Jun-2019]

  - User community: The users of the data marketplace are the data providers and the consumers.
  - Operator community: The operator community operates the data marketplace ecosystem. The operating community provides the computation and technical infrastructure for the data marketplace to function. The goal to have such a community avoid single ownerships.
- Open business architecture platform: The open business architecture platform is the platform i.e. the data marketplace itself which the community systems develop, operate and use. It describes the technical realization of a whole system for a data marketplace ecosystem. The main goal here is to provide a completely decentralized data marketplace ecosystem, which is open and flexible as far as possible but still provides nonfunctional requirements like safety, security, privacy and dependability.
- Relationship between the community system and open business architecture platform: The community system has various responsibilities which helps defining, development and the evolvement of the open business architecture platform. These responsibilities are the relationship between the community system and open business architecture platform. As the community uses and evolves the ecosystem understanding this relation is very important.

The proposed concept is community driven and proposes on the one side an initial concept for the community structure and on the other side an architecture which is open, flexible and secure and is aligned to this community. Nevertheless, this project is still work in progress and will be continuously expanded in the near future.

Although such a data ecosystem provides new opportunities for data consumers, providers and many other stakeholders, there are many challenges to be tackled. We identified various challenges in the data marketplace ecosystems. Although while writing this, our research is still work in progress we sketch some solutions for tackling few challenges.

- Open and secure infrastructure: In order to provide a fairness and transparency, new methods for secure but an open infrastructure is required.
- Metadata: In order to sell the data on the marketplace the seller needs to provide some description about the data. E.g. What is the data about, the size of the data etc. This information about the data is known as meta data. On one hand meta data gives the buyer the data information and is also used for search which helps the buyers to find relevant data. But providing such detailed relevant information about the data can be too much work for the sellers. One possibility for this challenge is Automatic generation of metadata. But the question still is how to generate the meta data automatically and protect it from unauthorized access.
- Data quality: Data is very different the any other commodity sold on electronic marketplaces. When a user buys a physical product online, this product can be returned or exchanged if he/she is unsatisfied with it. But the same does not applies for datasets. Once the buyer sees the datasets i.e. buys them, it cannot be returned. Thus, the biggest challenge we identify for trading data as a commodity is ensuring the data quality.

7.21 Jakob Rehof (TU Dortmund, DE)

When data ecosystems are understood to involve distributed networks of data providers and data consumers the question arises how to organize data logistics in the sense of a data supply network. Such issues may reasonably be attributed to the sub-field of data engineering.

A central data engineering challenge in this context is the automation of data supply networks, that is, essentially, the task of getting data from A to B in a form understood and desired by A and B. If two or more parties wanting to share data need to engage in arbitrarily complicated software engineering projects prior to making data flow among themselves, the prospect of realizing data ecosystems at large is fundamentally inhibited.

An interesting line of research motivated by these observations concerns the employment of component-oriented synthesis (much in the sense discussed at the previous Dagstuhl Seminar 14232 on Design and Synthesis from Components) as a means towards automating data supply chains. In particular, the automatic generation of complex data transformation functions and rendering functions appears to be a useful goal which should not be too far out of reach for practical exploitation. Based on some years of experience with research in type-based component-oriented synthesis in Dortmund, the idea of developing specific frameworks for use in data supply networks appears natural. The capacity of such frameworks to operate relative to given (but possibly dynamically changing) repositories of components seems useful here. For example, data provider A could inject data transformation functions based on domain specific knowledge of data semantics together with accompanying metadata. Data consumer B could inject domain specific transformers and rendering functions pertaining to the domain of use. An interesting fact about such a scenario is that the repository could evolve in a distributed fashion without any need for a central control other than the logical control implicit in overarching standards such as may be embodied in metadata formats. Thus a vision may arise which foresees distributed and (to a large extent) self-organizing networks of functionalities (the repository) accessible to automatic methods of code generation (synthesis) composing functions on demand to achieve a stated goal (for example, the transformation of data from A to B in a desired form). One can go further and imagine such a framework being combined with query languages and query mechanisms tailored for the network (such that, e.g., a data transformation function is automatically synthesized as a side-effect of executing a query against the network).

7.22 Simon Scerri (Fraunhofer IAIS – Sankt Augustin, DE)

Due to the well-understood AI opportunities presenting themselves in the last few years, there has been a steadily increasing and consistent interest in data sharing methods, solutions and practices. This has been observed internationally both at an industrial level, as well as at a political level. Seeing this as an opportunity to boost the data economy in Europe, the European Commission has reacted strongly by organising many events at attempting to invest in the convergence of technology and infrastructures that can enable the realisation and adoption of a pan-European data sharing space that can incorporate existing vertical, cross-sectoral, personal and industrial data spaces and enable broader participation. The realisation of an 'open' data sharing ecosystem that can serve the needs of all kinds of stakeholders also introduces exciting opportunities for scenarios that are not only restricted to B2B, but also enable data exchange possibilities for science, government and private citizens. For this vision to be achieved, the convergence of efforts at both industrial and socio-political level, through the adoption of standards that respect the existing legal and regulatory frameworks (e.g. GDPR for personal data), is being encouraged.

In view of this European vision for a data sharing space, the Big Data Value Association, an industry-driven international not–for-profit organisation with 200 European members (composed of large, small, and medium-sized industries as well as research and user organizations) has an activity group following and promoting advances in Data Sharing ecosystems.

Simon Scerri is one of the lead editors for the position paper published on the topic [20]. In its ambition to support the convergence of existing efforts, standards and technology in this sphere, the group is continuously seeking to collaborate with external experts and data practitioners. The position paper, for which a second version is planned for early 2020, includes a survey of the broad (international) technical landscape, and enlists the known opportunities (for business, science, government and public bodies and citizens), known challenges (legal compliance, technical, business and organisational, national and regional) and a list of key recommendations that can pave the way for the targeted convergence to materialise. The recommendations are addressed to both European policy makers and industry alike, as the two primary entities identified as having the highest likelihood of accelerating advances in the area.

The technical concerns and requirements discussed in the Dagstuhl seminar have helped to both confirm the challenges identified, as well as to bring to the focus additional concerns and relevant initiatives. In particular, the 'blueprints' for a high-level generic data sharing architecture that is open to all will be considered for a similar illustration in the second version of the BDVA position paper. For further information or updates, please contact Simon or refer to the BDVA Website Downloads section in the near future. In addition, an appeal is made for entities with an interest in data sharing practices to refer to a survey which is due to be published and promoted by the BDVA in January.

7.23 Julian Schütte (Fraunhofer AISEC – München, DE)

Future data ecosystems go beyond centralized cloud services and will require new technical paradigms and infrastructures to establish trust among participants. Central cloud providers currently serve not only as Identity Providers (IdP), but also as implicitly trusted data brokers which are expected to treat sensitive business data confidential, reliably enforce access constraints, and provide accounting and billing services. However, many upcoming business cases cannot be served by a single trusted cloud service – either due to legal reasons or due to technical necessities that require an orchestration of several services and stakeholders.

To overcome the limitations imposed by central cloud services, technical infrastructures that allow direct data exchange between participants without relying on central cloud services are needed. This raises various research questions, such as the establishment of trust between peers, the control of data flowing between services across enterprise boundaries, and ways to protect data while still being able to process it.

While some of these individual problems have been addressed by the research community in the past, synergies from recent developments in distributed architectures, in advanced cryptography, and in upcoming commodity hardware will allow to create a technical foundation for data-driven business cases that do not depend on centralized cloud providers. I would like to make the following contributions:

- Discussion of the main security building blocks needed to create trusted decentralized data ecosystems
- Insights into data usage control systems and their technical manifestation
- Presentation and discussion of mechanisms and upcoming standards for trust establishment, i.e. the foundations of automatically establishing trust in a remote party at a technical level

[20] http://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf

- Examples on how novel security mechanisms such as cryptographic access control in decentralized ledgers will foster new business cases in the logistics domain

7.24 Egbert Jan Sol (TNO – Eindhoven, NL)

The evolution of our (mechanical/electronics) industry towards a data driven industry (data eco-systems) involves many technologies. IoT data collection is one of the many aspects. It is the basis for coupling data to Digital Twin, AI-algorithm, but also data-driven businesses as servitization where measured usage of products leads to e.g. billing. IoT data collection faces several challenges to overcome. There are far too many different (fieldbus type) IoT data communication standards, limiting data use to local usage. But standardization as OPC-UA and usage of 5G will lead to an explosion of (large/big) data sets within and across businesses. Collected IoT data is not copy-right protected as it doesn't involved creative labor and needs extra legal and cyber security measurement. For AI usage data must be cleaned and for sending bills certain IoT data must be treated with DLT (distributed ledger tech/blockchain tech). And, maybe the biggest problem for business, is a huge lack of skilled, trained people with the proper digital skills as practically every non-university trained person above 35 years today didn't get any training in digital technologies 20 years ago at school when they were 15 years. With the rapid deployment of Industrie 4.0 mankind faces for the first time in history the need to life-long retrain every one in digital skills. For me the main question is what are the new big data/AI etc data technologies that are needed, that will be developed and how to make them usable, cq how can we educate and train people (and politicians) at academic, higher level and vocational level to understand them, to create the proper data-ecosystems and to deploy them. This challenge is not limited to my background in manufacturing industries, but is similar for food, medical and many other domains.

7.25 Maria-Esther Vidal (TIB – Hannover, DE)

During my academic and profesional work, I have been involved in diverse projects where the resolution of interoperability and quality issues across large volumes of heterogenoeus data sources is a pre-condition for effectively devising data integration. Albeit being projects from different areas, e.g., industry, biomedical, or scholarly communication, data variety and veracity seemed to be domain-agnostic even thought their resolution required domain specific knowledge. Motivated by this fact, the definion of generic frameworks able to identify interoperability issues while allowing for an effective data integration process has been one of my research topics and one of my interests in attending the seminar on Data Ecosystems. Currently, I am leading the tasks of integrating clinical data from electronic health records, gene sequences, and medical images, with open pharmacogenomics data and scientific publications. Natural language processing and semantics annotations from controlled vocabularies provide the basis for the fusion of these variety of data sources. During the seminar, I can contribute describing both challenges and solutions that we have tackled in the context of projects like iASiS [21] and BigMedilytics [22]. I can also discuss, the approaches that we have considered to address these issues during query processing.

[21] http://project-iasis.eu/
[22] https://www.bigmedilytics.eu/

References

1. Behzad Abdolmaleki, Karim Baghery, Helger Lipmaa, and Michał Zając. A subversion-resistant SNARK. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.
2. Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Data profiling: A tutorial. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1747–1751, 2017.
3. Ron Adner. Match your innovation strategy to your innovation ecosystem. *Harvard business review*, 84(4):98, 2006.
4. Dimitris Apostolou, Gregoris Mentzas, Bertin Klein, Andreas Abecker, and Wolfgang Maass. Interorganizational knowledge exchanges. *IEEE Intelligent Systems*, 23(4):65–74, 2008.
5. Marcelo Arenas, Pablo Barceló, Leonid Libkin, and Filip Murlak. *Foundations of Data Exchange*. Cambridge University Press, 2014.
6. Carlo Batini and Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications. Springer, 2016.
7. Eli Ben-Sasson, Iddo Bentov, Yinon Horesh, and Michael Riabzev. Scalable, transparent, and post-quantum secure computational integrity. *Eprint.Iacr.Org*, 2018.
8. Eli Ben-Sasson, Alessandro Chiesa, Eran Tromer, and Madars Virza. Succinct Non-Interactive Zero Knowledge for a von Neumann Architecture. *USENIX Security Symposium*, 2014.
9. Alexander Benlian, Daniel Hilkert, and Thomas Hess. How open is this platform? the meaning and measurement of platform openness from the complementers' perspective. *Journal of Information Technology*, 30(3):209–228, 2015.
10. Jan Bessai, Tzu-Chun Chen, Andrej Dudenhefner, Boris Düdder, Ugo de'Liguoro, and Jakob Rehof. Mixin composition synthesis based on intersection types. *Logical Methods in Computer Science*, 14(1), 2018.
11. Jan Bessai, Andrej Dudenhefner, Tzu Chun Chen, Ugo DE'LIGUORO, Jakob Rehof, et al. Mixin composition synthesis based on intersection types. In *TLCA 2015*, volume 38, pages 76–91. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.
12. Jan Bessai, Andrej Dudenhefner, Boris Düdder, Moritz Martens, and Jakob Rehof. Combinatory logic synthesizer. In *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pages 26–40. Springer, 2014.
13. Nitesh Bharosa, Marijn Janssen, Bram Klievink, and Yao-hua Tan. Developing multi-sided platforms for public-private information sharing. In Sehl Mellouli, Luis F. Luna-Reyes, and Jing Zhang, editors, *Proceedings of the 14th Annual International Conference on Digital Government Research - dg.o '13*, page 146, New York, New York, USA, 2013. ACM Press.
14. Vittorio Dal Bianco, Varvana Myllarniemi, Marko Komssi, and Mikko Raatikainen. The role of platform boundary resources in software ecosystems: A case study. In *2014 IEEE/IFIP Conference on Software Architecture*, pages 11–20. IEEE, 2014.
15. Kevin Boudreau. Open platform strategies and innovation: Granting access vs. devolving control. *Management science*, 56(10):1849–1872, 2010.
16. Sean Bowe, Alessandro Chiesa, Matthew Green, Ian Miers, Pratyush Mishra, Howard Wu, Cornell Tech, and Howard Wu. Zexe: Enabling Decentralized Private Computation. *Cryptology ePrint Archive*, 2018.
17. Georg Bramm, Mark Gall, and Julian Schütte. BDABE - blockchain-based distributed attribute based encryption. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018 - Volume 2: SECRYPT, Porto, Portugal, July 26-28, 2018*, pages 265–276, 2018.

18 Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. Bulletproofs: Short Proofs for Confidential Transactions and More. In *Proceedings - IEEE Symposium on Security and Privacy*, 2018.

19 Christian Burmeister, Dirk Lüttgens, and Frank T Piller. Business model innovation for industrie 4.0: why the "industrial internet" mandates a new perspective on innovation. *Die Unternehmung*, 70(2):124–152, 2016.

20 Diego Calvanese, Elio Damaggio, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. Semantic data integration in P2P systems. In *Databases, Information Systems, and Peer-to-Peer Computing, First International Workshop, DBISP2P, Berlin Germany, September 7-8, 2003, Revised Papers*, pages 77–90, 2003.

21 Ran Canetti, Yuval Ishai, Ravi Kumar, Michael K Reiter, Ronitt Rubinfeld, and Rebecca N Wright. Selective private function evaluation with applications to private statistics. In *Proceedings of the twentieth annual ACM symposium on Principles of distributed computing*, pages 293–304. ACM, 2001.

22 Michelle Cheatham, Isabel F. Cruz, Jérôme Euzenat, and Catia Pesquita. Special issue on ontology and linked data matching. *Semantic Web*, 8(2):183–184, 2017.

23 Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4), 2012.

24 Soon-Yong Choi, Dale O Stahl, and Andrew B Whinston. *The economics of electronic commerce*. Macmillan Technical Publ. Indianapolis, 1997.

25 Gianluca Cima. Preliminary results on ontology-based open data publishing. In *Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017*, 2017.

26 George Coker, Joshua Guttman, Peter Loscocco, Amy Herzog, Jonathan Millen, Brian O'Hanlon, John Ramsdell, Ariel Segall, Justin Sheehy, and Brian Sniffen. Principles of remote attestation. *International Journal of Information Security*, 10(2):63–81, 2011.

27 Diego Collarana, Mikhail Galkin, Ignacio Traverso-Ribón, Maria-Esther Vidal, Christoph Lange, and Sören Auer. MINTE: semantically integrating RDF graphs. In *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS*, 2017.

28 Brice Dattée, Oliver Alexy, and Erkko Autio. Maneuvering in poor visibility: How firms play the ecosystem game when uncertainty is high. *Academy of Management Journal*, 61(2):466–498, 2018.

29 Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati. Using ontologies for semantic data integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, pages 187–202. Springer, 2018.

30 Mark de Reuver, Bouke Nederstigt, and Marijn Janssen. Launch strategies for multi-sided data analytics platforms. In *ECIS 2018*, 2018.

31 Yuri Demchenko, Cees de Laat, and Peter Membrey. Defining architecture components of the big data ecosystem. In *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pages 104–112. IEEE, 2014.

32 Yuri Demchenko, Wouter Los, and Cees Laat. Data as economic goods: Definitions, properties, challenges, enabling technologies for future data markets. *ITUJournal - ICT Discoveries*, 1(2), 2018.

33 Charles Dhanaraj and Arvind Parkhe. Orchestrating innovation networks. *Academy of management review*, 31(3):659–669, 2006.

34 AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.

35 AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.

36 Boris Düdder and Omri Ross. Timber tracking: Reducing complexity of due diligence by using blockchain technology. *Available at SSRN 3015219*, 2017.

37 Jennie Duggan, Aaron J. Elmore, Michael Stonebraker, Magda Balazinska, Bill Howe, Jeremy Kepner, Sam Madden, David Maier, Tim Mattson, and Stan Zdonik. The BigDAWG Polystore System. *SIGMOD Rec.*, 44(2):11–16, August 2015.

38 Ben Eaton, Silvia Elaluf-Calderwood, Carsten Sørensen, and Youngjin Yoo. Distributed tuning of boundary resources: the case of apple's ios service system. *MIS Quarterly*, 39(1):217–243, 2015.

39 Benjamin Egelund-Müller, Martin Elsman, Fritz Henglein, and Omri Ross. Automated execution of financial contracts on blockchains. *Business & Information Systems Engineering*, 59(6):457–467, 2017.

40 Kemele M. Endris, Philipp D. Rohde, Maria-Esther Vidal, and Sören Auer. Ontario: Federated query processing against a semantic data lake. In *Database and Expert Systems Applications - 30th International Conference, DEXA 2019, Linz, Austria, August 26-29, 2019, Proceedings, Part I*, pages 379–395, 2019.

41 Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

42 Ronald Fagin, Laura M. Haas, Mauricio A. Hernández, Renée J. Miller, Lucian Popa, and Yannis Velegrakis. Clio: Schema mapping creation and data exchange. In *Conceptual Modeling: Foundations and Applications - Essays in Honor of John Mylopoulos*, pages 198–236, 2009.

43 Michael Franklin, Alon Halevy, and David Maier. From databases to dataspaces. *ACM SIGMOD Record*, 34(4):27–33, 2005.

44 Avigdor Gal. Uncertain schema matching. In *Encyclopedia of Big Data Technologies*. Springer, 2019.

45 Avigdor Gal, Haggai Roitman, and Roee Shraga. Heterogeneous data integration by learning to rerank schema matches. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018*, pages 959–964, 2018.

46 Günter Gans, Matthias Jarke, Stefanie Kethers, and Gerhard Lakemeyer. Continuous requirements management for organisation networks: a (dis)trust-based approach. *Requir. Eng.*, 8(1):4–22, 2003.

47 Günter Gans, Matthias Jarke, Gerhard Lakemeyer, and Dominik Schmitz. Deliberation in a metadata-based modeling and simulation environment for inter-organizational networks. *Information Systems*, 30(7):587–607, 2005.

48 Annabelle Gawer. Platform dynamics and strategies: from products to services. In Annabelle Gawer, editor, *Platforms, Markets and Innovation*, pages 45–77. Edward Elgar Publishing, 2009.

49 Annabelle Gawer. Bridging differing perspectives on technological platforms: Toward an integrative framework. *Research policy*, 43(7):1239–1249, 2014.

50 Craig Gentry. *A fully homomorphic encryption scheme*. PhD thesis, Stanford University Stanford, 2009.

51 Cheng Hian Goh, Stéphane Bressan, Stuart Madnick, and Michael Siegel. Context interchange: New features and formalisms for the intelligent integration of information. *ACM Trans. Inf. Syst.*, 17(3):270–293, 1999.

52 Behzad Golshan, Alon Y. Halevy, George A. Mihaila, and Wang-Chiew Tan. Data Integration: After the Teenage Years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 101–106, 2017.

53 Simon Gottschalk, Nicolas Tempelmeier, Günter Kniesel, Vasileios Iosifidis, Besnik Fetahu, and Elena Demidova. Simple-ml: Towards a framework for semantic data analytics workflows. In *Semantic Systems. The Power of AI and Knowledge Graphs - 15th International*

Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9-12, 2019, Proceedings, pages 359–366, 2019.

54 Rihan Hai, Sandra Geisler, and Christoph Quix. Constance: An intelligent data lake system. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 2097–2100, 2016.

55 Alon Y. Halevy, Anand Rajaraman, and Joann J. Ordille. Data Integration: The Teenage Years. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 9–16, 2006.

56 George Heineman, Armend Hoxha, Boris Düdder, and Jakob Rehof. Towards migrating object-oriented frameworks to enable synthesis of product line members. In *Proceedings of the 19th International Conference on Software Product Line*, pages 56–60. ACM, 2015.

57 O. Henfridsson and B. Bygstad. The generative mechanisms of digital infrastructure evolution. *MIS Quarterly: Management Information Systems*, 37(3):907–931, 2013.

58 Fritz Henglein and Jakob Rehof. Modal intersection types, two-level languages, and staged synthesis. In *Semantics, Logics, and Calculi - Essays Dedicated to Hanne Riis Nielson and Flemming Nielson on the Occasion of Their 60th Birthdays*, volume 9560 of *Lecture Notes in Computer Science*, pages 289–312, 2016.

59 Anne Immonen, Marko Palviainen, and Eila Ovaska. Requirements of an open data based business ecosystem. *IEEE Access*, 2:88–103, 2014.

60 Matthias Jarke. Data spaces: combining goal-driven and data-driven approaches in community decision and negotiation support. In *International Conference on Group Decision and Negotiation*, pages 3–14. Springer, 2017.

61 Matthias Jarke, Manfred Jeusfeld, and Christoph Quix. Data-centric intelligent information integration–from concepts to automation. *Journal of Intelligent Information Systems*, 43(3):437–462, 2014.

62 Thorhildur Jetzek, Michel Avital, and Niels Bjørn-Andersen. Generating sustainable value from open data in a sharing society. In *International Working Conference on Transfer and Diffusion of IT*, pages 62–82. Springer, 2014.

63 Manfred A. Jeusfeld, Matthias Jarke, and John Mylopouos. *Metamodeling for Method Engineering*. MIT Press, 2010.

64 Yasar Khan, Antoine Zimmermann, Alokkumar Jha, Vijay Gadepally, Mathieu D'Aquin, and Ratnesh Sahay. One size does not fit all: Querying web polystores. *IEEE Access*, 7:9598–9617, 01 2019.

65 Craig A. Knoblock and Pedro A. Szekely. Exploiting Semantics for Big Data Integration. *AI Magazine*, 36(1):25–38, 2015.

66 Phokion G. Kolaitis. Schema mappings, data exchange, and metadata management. In *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 61–75, 2005.

67 Phokion G. Kolaitis. Reflections on schema mappings, data exchange, and metadata management. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018*, pages 107–109, 2018.

68 Constantine E. Kontokosta. Energy disclosure, market behavior, and the building data ecosystem. *Annals of the New York Academy of Sciences*, 1295:34–43, 2013.

69 Sebastian Kortmann, Carsten Gelhard, Carsten Zimmermann, and Frank T Piller. Linking strategic flexibility and operational efficiency: The mediating role of ambidextrous operational capabilities. *Journal of Operations Management*, 32(7-8):475–490, 2014.

70 Sebastian Kortmann and Frank Piller. Open business models and closed-loop value chains: Redefining the firm-consumer relationship. *California Management Review*, 58(3):88–108, 2016.

71 Anastasia Krithara, Fotis Aisopos, Vassiliki Rentoumi, Anastasios Nentidis, Konstantinos Bougiatiotis, Maria-Esther Vidal, Ernestina Menasalvas, Alejandro Rodríguez González, Eleftherios Samaras, Peter Garrard, Maria Torrente, Mariano Provencio Pulla, Nikos Dimakopoulos, Rui Mauricio, Jordi Rambla De Argila, Gian Gaetano Tartaglia, and George Paliouras. iasis: Towards heterogeneous big data analysis for personalized medicine. In *32nd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2019, Cordoba, Spain, June 5-7, 2019*, pages 106–111, 2019.

72 Alice LaPlante and Ben Sharma. *Architecting data lakes: data management architectures for advanced business use cases*. O'Reilly Media, Sebastopol, 2016.

73 Kristin Lauter, Michael Naehrig, and Vinod Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the ACM Conference on Computer and Communications Security*, 2011.

74 Maurizio Lenzerini. Data Integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*, pages 233–246, 2002.

75 Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*, pages 233–246, 2002.

76 Maurizio Lenzerini. Managing data through the lens of an ontology. *AI Magazine*, 39(2):65–74, 2018.

77 Xavier Leroy. A formally verified compiler back-end. *Journal of Automated Reasoning*, 43(4):363, 2009.

78 Marten Lohstroh and Edward A Lee. An interface theory for the internet of things. In *SEFM 2015 Collocated Workshops*, pages 20–34. Springer, 2015.

79 Wolfgang Maass. *Elektronische Wissensmärkte: Handel von Information und Wissen über digitale Netze*. Springer-Verlag, 2009.

80 Wolfgang Maass, Wernher Behrendt, and Aldo Gangemi. Trading digital information goods based on semantic technologies. *Journal of Theoretical and Applied Electronic Commerce Research*, 2(3):18–35, 2007.

81 Majid Mohammadi, Amir Ahooye Atashin, Wout Hofman, and Yao-Hua Tan. Comparison of Ontology Alignment Systems Across Single Matching Task Via the McNemar's Test. *TKDD*, 12(4):51:1–51:18, 2018.

82 Corrado Moiso and Roberto Minerva. Towards a user-centric personal data ecosystem: The role of the bank of individuals' data. In Stuart Sharrock, editor, *2012 16th International Conference on Intelligence in Next Generation Networks (ICIN)*, pages 202–209, Piscataway, NJ, 2012. IEEE.

83 Michalis Mountantonakis and Yannis Tzitzikas. Large-scale semantic integration of linked data: A survey. *ACM Comput. Surv.*, 52(5):103:1–103:40, September 2019.

84 Kateryna Neulinger, Anna Hannemann, Ralf Klamma, and Matthias Jarke. A longitudinal study of community-oriented open source software development. In *International Conference on Advanced Information Systems Engineering*, pages 509–523. Springer, 2016.

85 Marcelo Iury S. Oliveira, Glória de Fátima A. Barros Lima, and Bernadette Farias Lóscio. Investigations into data ecosystems: a systematic mapping study. *Knowl. Inf. Syst.*, 61(2):589–630, 2019.

86 Jan Ondrus, Avinash Gannamaneni, and Kalle Lyytinen. The impact of openness on the market potential of multi-sided platforms: A case study of mobile payment platforms. *Journal of Information Technology*, 30(3):260–275, 2015.

87 B Otto, S Lohmann, S Auer, G Brost, J Cirullies, A Eitel, T Ernst, C Haas, M Huber, C Jung, et al. Reference architecture model for the industrial data space. *Fraunhofer-Gesellschaft, Munich*, 2017.

88 Boris Otto. Data ecosystems – conceptual foundations, constituents and recommendations for action.

89 Boris Otto and Matthias Jarke. Designing a multi-sided data platform: findings from the international data spaces case. *Electronic Markets*, 43(1):39, 2019.

90 Margherita Pagani. Digital business strategy and value creation: framing the dynamic cycle of control points. *Mis Quarterly*, pages 617–632, 2013.

91 Geoffrey Parker and Marshall Van Alstyne. Innovation, openness, and platform control. *Management Science*, 64(7):3015–3032, 2017.

92 Thomas F. J.-M. Pasquier and D. Eyers. Information flow audit for transparency and compliance in the handling of personal data. In *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)*, pages 112–117, April 2016.

93 Frank Piller. Digitale chancen und bedrohungen – geschäftsmodelle für industrie 4.0. In *VDI Statusreport*. VDI Verlag, 2016.

94 Christoph Pinkel, Carsten Binnig, Ernesto Jiménez-Ruiz, Evgeny Kharlamov, Andriy Nikolov, Andreas Schwarte, Christian Heupel, and Tim Kraska. Incmap: A journey towards ontology-based data integration. In *Datenbanksysteme für Business, Technologie und Web (BTW 2017), 17. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), 6.-10. März 2017, Stuttgart, Germany, Proceedings*, pages 145–164, 2017.

95 Michael E Porter and James E Heppelmann. How smart, connected products are transforming competition. *Harvard business review*, 92(11):64–88, 2014.

96 Christoph Quix, Rihan Hai, and Ivan Vatov. GEMMS: A generic and extensible metadata management system for data lakes. In *28th International Conference on Advanced Information Systems Engineering (CAiSE 2016)*, pages 129–136, 2016.

97 Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the European economic association*, 1(4):990–1029, 2003.

98 Jean-Charles Rochet and Jean Tirole. Two-sided markets: a progress report. *The RAND journal of economics*, 37(3):645–667, 2006.

99 Julian Schütte and Gerd Stefan Brost. A data usage control system using dynamic taint tracking. In *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, pages 909–916, March 2016.

100 Julian Schütte and Gerd Stefan Brost. LUCON: data flow control for message-based iot systems. In *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018*, pages 289–299, 2018.

101 Judith Gebauer Frank Farber Arie Segev. Internet-based electronic markets. *Electronic Markets*, 9(3):138–146, 1999.

102 Srinath Setty. Spartan : Efficient and general-purpose zkSNARKs without trusted setup. *Https://Eprint.Iacr.Org/2019/550*, 2019.

103 K. Su, J. Li, and H. Fu. Smart city and the applications. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 1028–1031, Sep. 2011.

104 B. Tan, S. L. Pan, X. Lu, and L. Huang. The role of is capabilities in the development of multi-sided platforms: the digital ecosystem strategy of alibaba.com. *Journal of the Association for Information systems*, 16(4):248–280, 2015.

105 Felix Ter Chian Tan, Barney Tan, and Shan L. Pan. Developing a leading digital multi-sided platform: Examining it affordances and competitive actions in alibaba.com. *Communications of the Association for Information Systems*, 38:738–760, 2016.

106 Nicolas Tempelmeier, Stefan Dietze, and Elena Demidova. Crosstown traffic - supervised prediction of impact of planned special events on urban traffic. *GeoInformatica. An International Journal on Advances of Computer Science for Geographic Information Systems*, 2019.

107 Nicolas Tempelmeier, Udo Feuerhake, Oskar Wage, and Elena Demidova. St-discovery: Data-driven discovery of structural dependencies in urban road networks. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2019, Chicago, IL, USA, November 5-8, 2019*, pages 488–491, 2019.

108 Nicolas Tempelmeier, Yannick Rietz, Iryna Lishchuk, Tina Kruegel, Olaf Mumm, Vanessa Miriam Carlow, Stefan Dietze, and Elena Demidova. Data4urbanmobility: Towards holistic data analytics for mobility applications in urban regions. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 137–145, 2019.

109 David Tilson, Kalle Lyytinen, and Carsten Sørensen. Research commentary—digital infrastructures: The missing is research agenda. *Information systems research*, 21(4):748–759, 2010.

110 Amrit Tiwana, Benn Konsynski, and Ashley A. Bush. Research commentary –platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information Systems Research*, 21(4):675–687, 2010.

111 Maria-Esther Vidal, Kemele M. Endris, Samaneh Jazashoori, Ahmad Sakor, and Ariam Rivas. Transforming heterogeneous data into knowledge for personalized treatments - A use case. *Datenbank-Spektrum*, 19(2):95–106, 2019.

112 Riad S. Wahby, Ioanna Tzialla, Abhi Shelat, Justin Thaler, and Michael Walfish. Doubly-Efficient zkSNARKs Without Trusted Setup. In *Proceedings - IEEE Symposium on Security and Privacy*, 2018.

113 Jonathan Wareham, Paul B Fox, and Josep Lluís Cano Giner. Technology ecosystem governance. *Organization Science*, 25(4):1195–1215, 2014.

114 Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

115 Youngjin Yoo, Ola Henfridsson, and Kalle Lyytinen. Research commentary—the new organizing logic of digital innovation: an agenda for information systems research. *Information systems research*, 21(4):724–735, 2010.

Participants

- Cinzia Cappiello
Polytechnic University of Milan, IT
- Ugo de' Liguoro
University of Turin, IT
- Yuri Demchenko
University of Amsterdam, NL
- Elena Demidova
Leibniz Universität Hannover, DE
- Boris Düdder
University of Copenhagen, DK
- Bernadette Farias Lóscio
Federal University of Pernambuco – Recife, BR
- Avigdor Gal
Technion – Israel Institute of Technology – Haifa, IL
- Sandra Geisler
Fraunhofer FIT – Sankt Augustin, DE
- Benjamin Heitmann
Fraunhofer FIT – Aachen, DE & RWTH Aachen, DE
- Fritz Henglein
Univ. of Copenhagen, DK & Deon Digital – Zürich, CH
- Matthias Jarke
RWTH Aachen, DE
- Jan Jürjens
Universität Koblenz-Landau, DE
- Maurizio Lenzerini
Sapienza University of Rome, IT
- Wolfgang Maaß
Universität des Saarlandes – Saarbrücken, DE
- Paolo Missier
Newcastle University, GB
- Boris Otto
Fraunhofer ISST – Dortmund, DE & TU Dortmund, DE
- Elda Paja
IT University of Copenhagen, DK
- Barbara Pernici
Polytechnic University of Milan, IT
- Frank Piller
RWTH Aachen, DE
- Andreas Rausch
TU Clausthal, DE
- Jakob Rehof
TU Dortmund, DE
- Simon Scerri
Fraunhofer IAIS – Sankt Augustin, DE
- Julian Schütte
Fraunhofer AISEC – München, DE
- Egbert Jan Sol
TNO – Eindhoven, NL
- Gerald Spindler
Georg August Universität – Göttingen, DE
- Maria-Esther Vidal
TIB – Hannover, DE