Chapter 2
Enabling Knowledge Flows in an Intelligent Systems Data Ecosystem

Edward Curry and Adeboyega Ojo

Keywords Intelligent systems · Data ecosystem · Open systems · Digital twins · Internet of Things · Data platforms · Smart environment

2.1 Introduction

In data ecosystems, vast amounts of data move among actors within complex information supply chains that can form in different ways around an organisation, community technology platforms, and within or across sectors. This chapter explores the role a data ecosystem can play in the design of intelligent systems to support data-rich Internet of Things (IoT)-based smart environments. The chapter examines different elements of an intelligent systems data ecosystem that are critical to understanding the data management and sharing challenges they present.

In Sect. 2.2, we establish the foundations of an intelligent systems data ecosystem and explore the increasing role data is playing in the design of intelligent systems. Section 2.3 details the challenge to support the exchange of knowledge within open systems in dynamic environments, with Sect. 2.4 outlining the Knowledge Value Ecosystem (KVE) Framework to support knowledge sharing. Sections 2.5, 2.6, and 2.7 explain the framework in more detail and how knowledge, value, and ecosystem barriers are overcome. A pay-as-you-go iterative boundary crossing process to overcome these barriers is discussed in Sect. 2.8. Section 2.9 details the requirements for data platforms to support the sharing of data between intelligent systems within Internet of Things-based smart environments and a summary is provided in Sect. 2.10.

2.2 Foundations

As we begin the third decade of the twenty-first century, we are at the beginning of a great wave of convergence of enabling technologies from the Internet of Things (IoT), 5G, high-performance computing, and edge computing to big data, cloud

E. Curry, *Real-time Linked Dataspaces*,
https://doi.org/10.1007/978-3-030-29665-0_2

computing, and Artificial Intelligence (AI). Smart environments are generating significant quantities of data from digital infrastructure that is driving a new wave of data-driven intelligent systems. Over the last decade, the term "Big Data" was used by different major players to label data with different attributes. The first definition, by Doug Laney of META Group (later acquired by Gartner) defined big data using a three-dimensional perspective: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision-making, insight discovery and process optimization" [20]. Loukides [21] defines big data as "when the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam." Jacobs [22] describes big data as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time". Big data brings together a set of data management challenges for working with data under new scales of size and complexity. Many of these challenges are not new. What is new, however, are the challenges raised by the specific characteristics of big data related to the three V's:

- *Volume (amount of data)*: Dealing with large scales of data within data processing (e.g. healthcare and logistics)
- *Velocity (speed of data)*: Dealing with streams of high-frequency incoming real-time data (e.g. sensors and IoT devices)
- *Variety (range of data types/sources)*: Dealing with data using differing syntactic formats (e.g. spreadsheets, XML, DBMS), schemas, and semantic meanings (e.g. Enterprise Data Integration).

The V's of big data challenge the fundamentals of existing technical approaches and require new forms of data processing to enable enhanced decision-making, insight discovery, and process optimisation. As the big data field matured, other V's have been added, such as Veracity (documenting quality and uncertainty) and Value. The value of data within a smart environment can be considered in the context of the dynamics of knowledge-based organisations [23], where the processes of decision-making and organisational action are dependent on the process of sense-making and knowledge creation.

Through the generation and analysis of data from the smart environment, data-driven systems are transforming our everyday world. From the digitisation of traditional infrastructure (smart energy, water, and mobility), the revolution of industrial sectors (smart autonomous cyber-physical systems, autonomous vehicles, and Industry 4.0), to changes in how our society operates (smart government and cities). At the other end of the scale, we see more human-centric thinking in our systems where users have growing expectations for highly personalised digital services for the "Market of One".

The digital transformation is creating an ecosystem with data on every aspect of our world spread across a range of information systems. Data ecosystems present new challenges to the design of intelligent systems that require a reconsideration of how we deal with the data management needs of large-scale, data-rich smart environments. Intelligent systems need to support openness, flexibility, and dynamicity [24] with the ability to deal with incremental change at minimum cost.

To understand the emerging data management challenges, we explore the design of intelligent systems within smart environments and the need to support knowledge flows within data ecosystems.

2.2.1 Intelligent Systems Data Ecosystem

Within a data ecosystem, participants (individual or organisation) can create new value that no single participant could achieve by itself [25]. A data ecosystem can form in different ways, around an organisation, a community of interest (music), a geographical location (city), or within or across industrial sectors (manufacturing, pharmaceutical). In the context of a smart environment, the data ecosystem metaphor is useful to understand the challenges faced with the cross-fertilisation and exchange of knowledge from different intelligent systems within the environment.

A key challenge within the design of intelligent systems is the need to extract valid and accurate insights from the data generated by a smart environment to make useful and meaningful decisions for business and society. Figure 2.1 details the data ecosystem for a connected autonomous vehicle where a community of interacting information systems share and combine their data to provide a holistic functional view of the car, passengers, city mobility, and service and infrastructure providers. Data may be shared about the current operating conditions of the vehicle, traffic flows, or context of the passengers (e.g. a family on holiday or a business executive moving between meetings) to support real-time decision-making, personalised digital services, or data on past observations to improve learning processes.

An intelligent systems data ecosystem (see Fig. 2.2) describes a community of interacting information systems that can share and combine their data to provide a

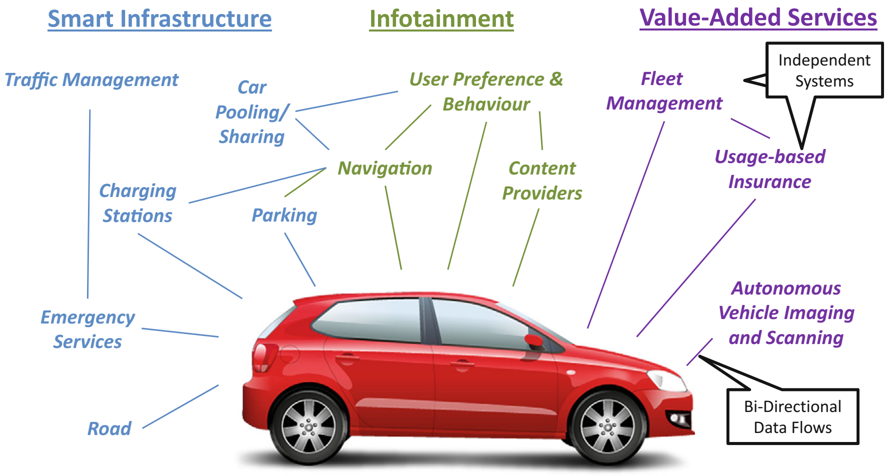


Fig. 2.1 Connected and autonomous vehicle data ecosystem [1]

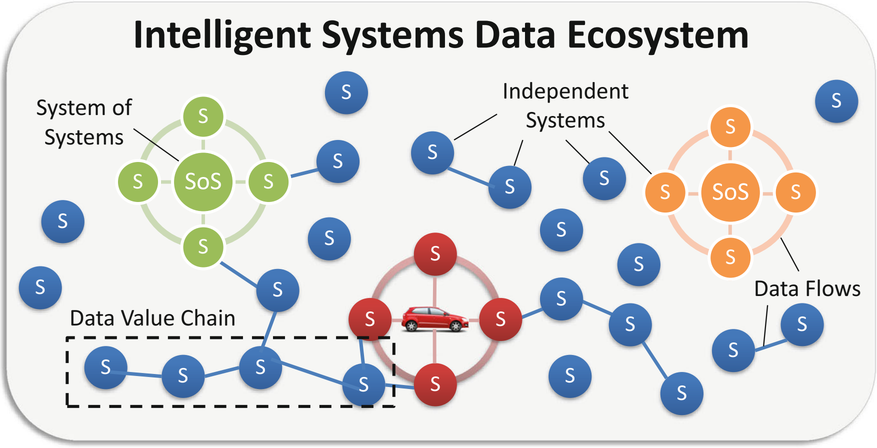


Fig. 2.2 An intelligent systems data ecosystem

functional view of the environment [1]. The ecosystem supports the flow of data among systems, enabling the creation of data value chains to understand, optimise, and reinvent processes that deliver insight to optimise the overall environment. In a data value chain, information flow is described as a series of steps needed to generate value and useful insights from data [15]. Systems within the ecosystem can also come together to form a System of Systems.

2.2.2 System of Systems

The need for multiple intelligent systems within a smart environment to work together is becoming a standard requirement. Sharing data among intelligent systems is critical if we are to extract the maximum value from IoT-based smart environments. Smart cities are showing how different systems within the city (e.g. energy and transport) can collaborate to maximise the potential to optimise overall city operations [26]. Digital services are expected to deliver a personalised and seamless user experience by bringing together relevant user data from multiple systems [16]. Building these systems requires a System of Systems (SoS) approach to connect systems that cross organisational boundaries, come from various domains, (e.g. finance, manufacturing, facilities, IT, water, traffic, and waste) and operate at different levels (e.g. region, district, neighbourhood, building, business function, individual). The joint ISO/IEC/IEEE definition of an SoS is that it "brings together a set of systems for a task that none of the systems can accomplish on its own. Each constituent system keeps its management, goals, and resources while coordinating within the SoS and adapting to meet SoS goals" [27]. Maier [28] identified a set of characteristics to describe an SoS:

- *Operational Independence*: Constituent systems can operate independently from the SoS and other systems.
- *Managerial Independence*: Different entities manage the constituent systems.
- *Geographic Distribution*: Is the degree to which a system is widely spread or localised.
- *Evolutionary Development*: The SoS, and its behaviour evolves, requiring changes to system interfaces to be maintained and kept consistent.
- *Emergent Behaviour*: New emergent behaviour can be observed when the SoS changes.

There are many systems engineering challenges in bringing together the constituent systems into an SoS at the data, service, process, and organisational levels. Many of the above characteristics of an SoS give us insights into the knowledge, value, and ecosystem boundaries that exist in bringing an SoS together, and the different types of interests possible at management and operational levels of systems. At the data level, intelligent systems can benefit from leveraging data from the availability of large volumes and variety of data and streams in the smart environment, which can be used to fuel intelligent, evidence-based decision-making.

2.2.3 From Deterministic to Probabilistic Decisions in Intelligent Systems

When it comes to making decisions in intelligent systems, there are two general approaches: deterministic (model-driven) and probabilistic (data-driven). A critical difference between the approaches can be explored by considering the costs and level of reliability and adaptability they provide within intelligent systems. There is a tension between reliability, predictability, and cost [29]: usually the more dependable and reliable the intelligent system needs to be, the more cost is associated with its development. Typically, we can see deterministic systems as reliable but with high costs to develop and adapt, and probabilistic as low cost to build and adapt, but less reliable. Take the example of the autonomous connected car, where we have the strict requirements of safety-critical autonomous driving systems (where a failure may lead to loss of life or serious personal injury) to the "good enough" requirements of the infotainment systems (where a failure is acceptable and merely an inconvenience to the user).

Within early smart environments, the level of data available was limited due to the high cost of digitisation. Sensors were expensive to purchase and install, resulting in the prudent use of resources. Conventional intelligent systems typically targeted "high-value" opportunities where the cost savings and benefits could justify the high cost of investment needed. Often these would be safety- or mission-critical systems that required higher levels of reliability. Due to the lack of sensor data and the need for high levels of reliability, deterministic approaches were an obvious choice for "conventional" intelligent systems. In this approach, the environment is optimised based on a formal deterministic model where a set of rules and/or equations detail the decision logic for the intelligent system that is used to control the activity in the environment efficiently and predictably. Adapting the intelligent system to meet changes in the environment is a costly process as the model, and its rules, need to be updated by expert systems engineers.

In the probabilistic approach, the core of the decision process is a statistical model that has been learnt from an analysis of "training" data to "discover" the structure of a decision model automatically from the observed data (e.g. driver behaviour). Thus, a fundamental requirement of data-driven approaches is the need for data to fuel the training of the algorithms. A lack of data, and training data, within a smart environment has limited the use of data-driven approaches.

As the IoT is enabling the deployment of lower-cost sensors, we are seeing more extensive adoption of IoT devices/sensors and gaining more visibility (and data) into smart environments. Smart environments are generating different types of data with an increase in the number of multimedia devices deployed, such as vehicle and traffic cameras. The emergence of the Internet of Multimedia Things (IoMT) is resulting in large quantities of high-volume and high-velocity multimedia event streams that need to be processed [30]. The result is a data-rich ecosystem of structured and unstructured data (e.g. images, video, audio, and text) detailing the smart environment that can be exploited by data-driven techniques. It is estimated that a single connected car will upload about 25 gigabytes of data per hour, while a vehicle fitted with an Autonomous Vehicle Imaging and Scanning system generates and processes about 4 TB of data for every autonomous driving hour (https://www.datamakespossible.com/evolution-autonomous-vehicle-ecosystem/).

The increased availability of data has opened the door for the use of data-driven probabilistic models, and their use within smart environments is becoming more and more commonplace for "good enough" scenarios. As a result, the conventional rule-based approach is now being augmented with data-driven approaches that support optimisations driven by machine learning, cognitive and AI techniques that are opening new possibilities for the design of intelligent systems. For example, pedestrian detection is challenging to implement in a rule-based approach. However, deep learning models for object detection and semantic segmentation using a dash-mounted camera are highly effective at detecting pedestrians.

Intelligent systems can now adapt to changes in the environment by leveraging the data generated in the environment within their learning process to improve performance. If intelligent systems share data on their operational experiences, a pool of data can be created to improve the overall learning processes of all the systems, a form of collective AI through the "wisdom of the systems". Because the process is data-driven, it can be run and re-run at low cost. This critical role of data in enabling adaptability and collective machine intelligence makes it a valuable resource.

Within the context of smart environments, data-driven approaches have been used to optimise the operation of infrastructure, such as the energy and transportation systems [31]. However, the adoption of data-driven approaches is about to increase significantly across a range of industries and sectors with the use of Digital Twins.

2.2.4 Digital Twins

Within the business community, the metaphor of a “Digital Twin” is gaining popularity as a way to explain the potential of IoT-enabled assets and smart environments [32]. A digital twin refers to a digital replica of a physical asset (car), processes (value chain), system (transport), or physical environment (building). As illustrated in Fig. 2.3, the digital twin provides a digital representation (i.e. simulation model or data-driven model) that updates and changes as the “physical twin” changes. The digital representation provided by the digital twin can be analysed to optimise the operation of the physical twin.

Digital twins are constructed from multiple sources of data, including real-time IoT sensors, historical sensor data, traditional information systems, and human input from domain and industrial experts. With the use of advanced analytics, machine learning, and AI techniques, the digital twin can learn the optimal operating conditions of the physical twin and optimise the physical twins’ operations in areas such as performance, maintenance, and user experience. One of the most promising outputs from a digital twin analysis is the possibility to find root causes of potential anomalies which can happen (prediction) and improve the physical process (innovation).

Digital twins can range from human organs such as the heart and lungs to aircraft engines and city-scale twins. For example, the SmartSantander smart city project has deployed tens of thousands of IoT-connected sensor devices in large cities across

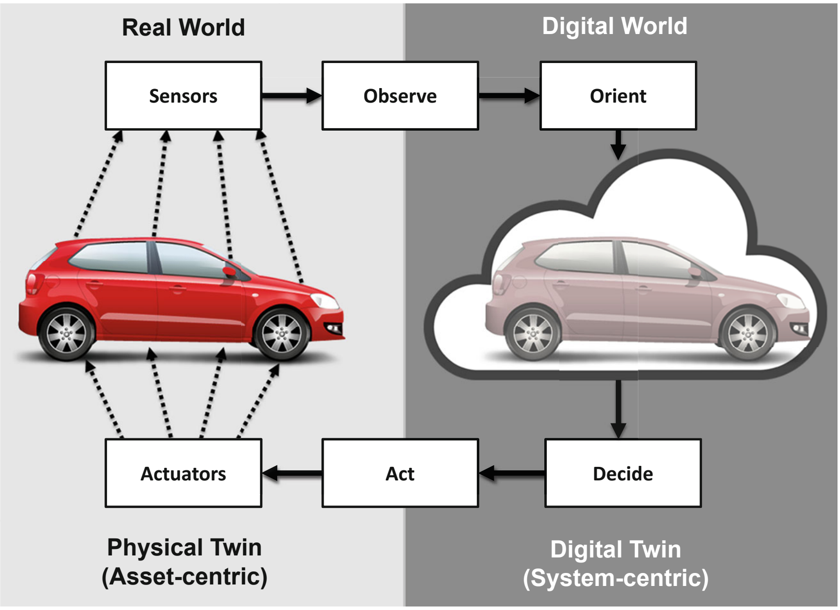


Fig. 2.3 Information flow and processing steps within a digital twin

Europe [33]. The sensing capabilities of these devices are wide-ranging, including solar radiation, wind speed and direction, temperature, water flow, noise, traffic, public transport, rainfall, and parking. The devices provide a digital representation of the city, which enables visibility into city processes and operations to support analysis and optimisation.

Datafication is creating an ecosystem of data on every aspect of our world spread across a range of information systems. In order for digital twins and intelligent systems to maximise the benefits from the resulting data ecosystems, we need to rethink how we exchange knowledge among open intelligent systems in dynamic environments.

2.3 Knowledge Exchange Between Open Intelligent Systems in Dynamic Environments

The design of intelligent systems, especially ones enabled by IoT, has to accommodate the needs of dynamic environments, where system participants continuously join and leave the environment. Vermesan et al. call this phenomenon "fluid systems" that are continuously changing and adapting, "in IoT systems it is very common to have nodes that join and leave the network spontaneously" [34]. This dynamic nature puts constraints on the assumptions that can be made within the design of intelligent systems and the assumption of having full understanding or control over the systems in the environment. This has led to the need for "open" intelligent systems which can adapt to their environment and learn from its interaction with the changing dynamics of the environment and the different systems operating within it.

While the term "open" has been frequently used in the literature to describe large-scale distributed systems, for example, Ciliaet et al. [35], a broad consensus has not been reached on its definition. Looking to the early works in system theory, we can draw upon the definition commonly used in this field as a system that has external interactions in the form of information, energy, or matter transfer through the system boundary [36]. A boundary here separates the system from its environment. For example, in biology, a cell exchanges chemicals with its environment through its membrane, and thus, it is an open system from this perspective.

The concept of boundary objects is established in the literature on knowledge sharing and reuse. Boundary objects are used to understand and coordinate the interactions among actors with varying information and knowledge needs to establish a shared point of reference during interactions [37]. Carlile formulates suggestions for managing knowledge across boundaries and provides the 3-T framework for knowledge exchange across system boundaries within the area of organisation science [37]. While Carlile's framework focuses on the exchange of knowledge between product development teams (the "systems" in this case), its foundations can be traced back to the Shannon-Weaver model with implications for information systems. The 3-T framework defined the task of knowledge exchange as a task of crossing three boundaries among systems: syntactic, semantic, and pragmatic. We interpret these boundaries from Carlile for knowledge exchange among open systems in dynamic environments and extend them with new boundaries for value exchange and ecosystem coordination.

In Fig. 2.4, the inverted pyramidal frustum shape shows the spectrum of tasks between well-known and novel tasks that need to be undertaken to exchange knowledge within an ecosystem. Systems A and B interact within this spectrum with correspondence to boundary objects that exist at three levels:

- *Knowledge Boundaries*: Exist where differences and dependencies among systems exist at the semantic and the administrative levels. A common lexicon needs to be developed to transfer and assess knowledge among systems in the classical sense from Shannon [38]. However, as Shannon noted, a common lexicon may not always be sufficient to share knowledge among systems. Distinct systems will have differences and dependencies that are unclear with multiple possible interpretations which create a semantic boundary to knowledge sharing. The administrative boundary describes how close or far in terms of control are the systems. A close control means that many assumptions can hold concerning data management guarantees (e.g. data consistency, availability, and quality), while a far control refers to weaker or no guarantees. To cross the knowledge boundary, it is necessary to develop common meanings to provide a means of sharing and assessing knowledge at a boundary. This requires new agreements on the translation of each system to the commonly shared meaning and an agreed upon protocol for access.
- *Value Boundaries*: Systems generally serve the interest of their participants, with different systems serving the different interests of their users. Cultural, organisational, and social interests can impede the sharing of knowledge among systems. To overcome the value boundary, it is necessary to develop common

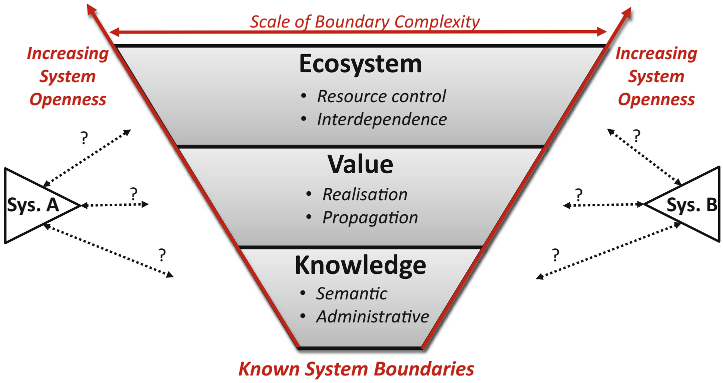


Fig. 2.4 Knowledge exchange between two systems within an ecosystem. Based on concepts from the 3-T Framework [37]

interests among systems, and their participants, to provide sufficient motivation for knowledge sharing. The value transformation necessary to create common interests requires significant practical and political effort, and the value must be propagated within the ecosystem.

- *Ecosystem Boundaries*: Ecosystems generally have different levels of interdependence between systems in both the technical and management sense. The ecosystem can create the conditions for a marketplace competition among participants or enable collaboration among diverse, interconnected participants that depend on each other for their mutual benefit. Another key factor is the control of key data resources within the ecosystem. Who own the key data resources? Is the data available to all participants in the ecosystem? Are there commercial terms of use? A close ecosystem coordination framework would provide clear answers to these questions, while loose coordination means less predictability on the behaviour of participants within the data ecosystem. To overcome ecosystem boundaries, it is necessary to understand and support the social, political, organisational, and business changes needed for ecosystem coordination.

The more open, distributed, and heterogeneous the environment becomes, the more significant these boundaries become, especially the latter ones where openness may introduce more novelty and uncertainty. Crossing boundaries requires mutual agreements among participants, which implies cost. The need for mutual agreements among participants adds to the technical issues an essential social dimension. Overcoming the differences among systems generates costs to the systems involved where domain-specific knowledge, as well as the common knowledge used, may need to be transformed to share and assess knowledge among the systems effectively.

There is an inherent need to design intelligent systems with the ability to scale and cross system boundaries. To effectively cross the ecosystem boundary for multiple systems within an open environment, each system must be able to represent current and more novel forms of knowledge, learn about their consequences, and transform their domain-specific knowledge accordingly. Intelligent systems within dynamic environments need to support the “social” agreement needed to share knowledge among them. We capture the capabilities needed to overcome knowledge sharing barriers among intelligent systems in the KVE Framework.

2.4 Knowledge Value Ecosystem (KVE) Framework

In order to cross the three boundaries of sharing knowledge among open intelligent systems in dynamic environments, we propose the Knowledge Value Ecosystem (KVE) Framework (Fig. 2.5). The KVE Framework, an extension of the 3-Ts Framework, tackles each boundary using the following capabilities:

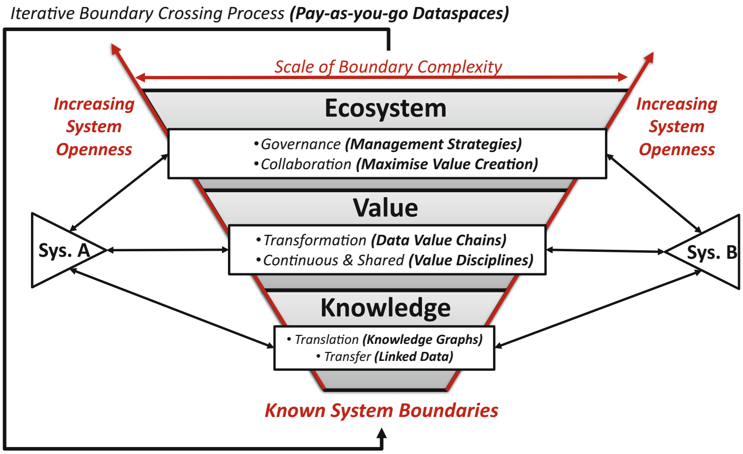

Iterative Boundary Crossing Process (Pay-as-you-go Dataspaces)
Scale of Boundary Complexity
Increasing System Openness
Increasing System Openness
Ecosystem
•Governance (Management Strategies)
•Collaboration (Maximise Value Creation)
Value
•Transformation (Data Value Chains)
•Continuous & Shared (Value Disciplines)
Knowledge
• Translation (Knowledge Graphs)
• Transfer (Linked Data)
Sys. A
Sys. B
Known System Boundaries

Fig. 2.5 The Knowledge Value Ecosystem (KVE) Framework for enabling knowledge flows within an ecosystem

- *Knowledge Transfer and Translation*: Knowledge boundaries are crossed by tackling administrative barriers using capabilities for the transfer of data using open web standards for data publishing and sharing knowledge as linked data, while semantic boundaries are crossed with capabilities to establishing common meanings among intelligent systems using knowledge graphs (see Sect. 2.5.3). Together, linked data and knowledge graphs can be used to support an incremental approach to reaching agreements on the transfer and translation of meaning among multiple intelligent systems.
- *Continuous and Shared Value Transformation*: Value boundaries are crossed with capabilities for transforming the interests of individual participants into common interests within new shared data value chains. The shared data value chain approach can provide a clear value proposition to support the political effort necessary from both a business case and an organisational perspective. It is important that the value created is continuously shared between participants along the value chain to motivate their contribution and support the sustainability of the data ecosystem.
- *Ecosystem Governance and Collaboration*: The nature of the ecosystem, the participants, and their dynamics will affect the management strategies needed to support the social, political, and organisational changes needed. Within a well-functioning data ecosystem, the participants are efficiently and effectively collaborating to exchange knowledge to maximise value creation.
- *Iterative Pay-As-You-Go Process*: Typically, the process of crossing boundaries, especially ecosystem boundaries, cannot be resolved with a single attempt. It requires an *Iterative Boundary Crossing Process* which supports trial and error in transforming complex knowledge across system boundaries. An iterative approach can support a learning process to improve the boundary crossing

Table 2.1 Boundaries, barriers, and capabilities within the KVE Framework and proposed implementation within a data platform

Boundary	Barrier	Capability	Data platform implementation
Knowledge	Administrative	Transfer	Linked data
	Semantic	Translation	Knowledge graphs
Value	Realisation	Transformation	Data value chains
	Propagation	Continuous and shared	Value disciplines for data networks
Ecosystem	Resource control	Governance	Management strategies
	Interdependence	Collaboration	Maximise value creation
		Iterative boundary process	Pay-as-you-go dataspaces

capability. Dataspaces provide a pay-as-you-go approach to support incremental data management within the ecosystem.

Within this approach, there is natural support between the different capabilities as linked data can support the definition and sharing of knowledge graphs, which can then both support the creation of value chains in a data ecosystem and motivate the need for collaboration between participants. The first capabilities (knowledge transfer and translation) can be captured in the technical design of a data platform. The second (continuous and shared value transformation) requires a higher-level value transformation among systems together with a cultural transformation of the stakeholders to promote data sharing and creation of new data value chains among systems. The purpose of the third capability (governance and collaboration) is to gradually improve the overall operation of the ecosystem to maximise benefits for all participants. Finally, the iterative boundary crossing process can support all the capabilities to improve over time following a pay-as-you-go approach. The alignment of the boundaries, barriers, capabilities, and their implementation is detailed in Table 2.1. We will now introduce each of these capabilities in more detail and explore how they can be implemented within the design of a data platform to support knowledge sharing among intelligent systems in a data ecosystem.

2.5 Knowledge: Transfer and Translation

In order to cross the knowledge boundaries of systems, two capabilities are needed: transfer and translation. Within the KVE Framework, knowledge boundaries are crossed by using an entity-centric model that establishes common meanings among systems using knowledge graphs expressed using linked data.

2.5.1 Entity-Centric Data Integration

Data integration projects typically focus on one-off point-to-point integration solutions among two or more systems in a customised but inflexible and non-reusable manner—this limits both the information flow and its oversight among systems to those that have been integrated. Entity-centric data integration takes a different form to traditional schema-level integration within the relational model. The entity-centric data integration model is based on global identifiers representing objects or concepts that can be reused or reconciled among different datasets (or systems).

Entity-centric data integration facilitates the co-existence of different perspectives and points of view of entities and a decentralised evolution of the data. At the same time, the use of linked data vocabularies, and the specification of conceptual models for a domain under the Resource Description Framework (RDF) model are used to facilitate the interoperability and semantic integration among different datasets for specific domains. This entity-centric integration of knowledge graphs using linked data has a number of virtues to represent large, complex, and heterogeneous conceptual models as detailed by [39–41]:

- *Support for the representation of sparse data*: RDF(S) is based on a graph data model, which supports a sparse data model.
- *Schema flexibility*: RDF(S) datasets are schema-less and can evolve in a decentralised manner.
- *Represent and map to/from other data models*: Data in a relational or in other formats (e.g. CSV) can be represented and systematically mapped to RDF [42].

These characteristics make entity-centric knowledge graphs an ideal approach for establishing a shared meaning among systems to cross knowledge boundaries. When knowledge graphs are expressed using linked data, they can be created in a fashion that allows two systems to be easily linked to each other on the information-level (data) not the infrastructure-level (system) by focusing more on the conceptual similarities (shared understanding). The combination of knowledge graphs and linked data meets many of the FAIR data principles for data management (see Sect. 2.8), including persistent identifiers, metadata, and open protocols. The approach provides a means for translating knowledge across the knowledge boundaries among systems. It allows separate systems that were designed independently to be later joined/linked at the edges, for interoperability to be added incrementally when needed and where cost-effective, and for the meaning of data to be expressed in a mixture of vocabularies.

2.5.2 Linked Data

In order to cross the administrative boundaries of systems to support data transfer, we propose the use of linked data. Linked data leverages open protocols and W3C standards of the web architecture for sharing structured data on the web. The fundamental concept of linked data is that data is created with the mindset that it will be shared and reused by others. The objective is to expose the data within existing systems but only link the data when it needs to be shared. Linked data provides a decentralised incremental approach for information sharing based on the creation of a global information space [43]. Linked data has the following characteristics:

- *Open*: Linked data is accessible through a variety of applications because it is expressed in open, non-proprietary syntactic format.
- *Modular*: Linked data can be combined (mashed-up) with any other pieces of linked data. No planning is required to integrate two data sources if they both use linked data standards.
- *Scalable*: It is easy to add and connect more linked data to existing linked data, even when the terms and definitions that are used change over time.

Linked data uses standards, tools, and techniques from work on the semantic web to facilitate sharing and reuse of data across domains. It primarily uses a graph-based representation framework for structuring data and uses standard ontology languages for defining the semantics of data. Ontologies (or vocabularies) provide a shared understanding of concepts and entities within a domain of knowledge which supports automated processing for data using semantic web tools. Thus, the use of linked data at the syntactic level can support the establishment of a common lexicon. At the semantic level, it can also support the establishment of shared meanings.

Linked data when used together with the dataspace approach provides a framework for a decentralised pay-as-you-go data integration with a standardised data model representation providing a minimum level of integration and where Universal Resource Identifiers (URIs) and the Domain Name Systems (DNS) provide a global-level identification scheme, which facilitates the referencing of data entities among different datasets. The RDF standard provides a common interoperable format and model for data linking and sharing on the web. RDF is the basic machine-readable representational format used to represent information. It is a general method for encoding graph-based data that is self-describing, meaning that the labels of the graph describe the data itself.

Linked data uses web standards in conjunction with four basic principles for exposing, sharing, and connecting data. These principles are:

- *Naming*: Use of URIs as names to identify things such as a person, a building, a device, an organisation, an event or even concepts such as risk exposure or energy and water consumption, simplifies reuse and the integration of data.
- *Access*: Use of URIs based on HyperText Transfer Protocol (HTTP) so that people can look up those names—URIs are used to retrieve data about objects using standard web protocols. For an employee, this could be their organisation and job classification, for an event, this may be its location time and attendance, for a device, this may be its specification, availability, and price.

- *Format*: When a URI is looked up (dereferenced) to retrieve data, it provides useful information using a standardised format, ideally, in web standard formats such as RDF.
- *Contextualisation*: Include links to other URIs so that more information can be discovered. Retrieved data may link to other data, thus creating a data network; for example, data about a product may link to all the components it is made of, which may link to their supplier.

Using these technologies, we can support data transfer among intelligent systems by using: (1) URIs to name things; (2) RDF for representing data; (3) Linked data principles for publishing, linking, and integration; (4) Vocabularies to establish and share understanding; and (5) Bottom-up incremental agreement.

2.5.3 Knowledge Graphs

Overcoming semantic boundaries among systems requires a common understanding of meaning among systems for knowledge to be shared. Within the KVE Framework, semantic boundaries are crossed by establishing common meanings among systems using knowledge graphs expressed using linked data. Knowledge graphs and linked data can be used to support an incremental approach to reaching agreements on the translation of the meaning of knowledge among systems.

In 2012 Google coined the term “Knowledge Graph” to refer to their use of information gathered from multiple sources to enrich their services, including search engine results. The term has also been used to refer to Semantic Web knowledge bases such as DBpedia or YAGO. As defined by Paulheim [44] a “knowledge graph (1) mainly describes real-world entities and their interrelations, organised in a graph, (2) defines possible classes and relations of entities in a schema, (3) allows for potentially interrelating arbitrary entities with each other and (4) covers various topical domains.” As illustrated in Fig. 2.6, a knowledge graph is just a set of entities (e.g. Marie Curie and France), a set of relations between those entities’ (e.g. “knownFor” and “wasResidentOf”), and a set of facts (see Table 2.2). Facts are the combination of the entities and relationships “Marie Curie, wasResidentOf, France”. More formally, a knowledge graph is a tuple (E, R, G), where:

- E is a set of nodes, each representing an **entity** in the domain.
- R is a set of edge labels, each representing a **predicate**, or a semantic relation type.
- $G \subseteq E \times R \times E$ is a set of ⟨subject, predicate, object⟩ **triples**, denoting **facts**.

Knowledge graphs provide a flexible knowledge representation structure that can describe entities and concepts that may come from multiple systems and domains, and at varying levels of granularity. Knowledge graphs can be used to create large knowledge bases (see Table 2.3). However, managing graphs of these sizes poses several challenges regarding quality, coherence, performance, and interaction.

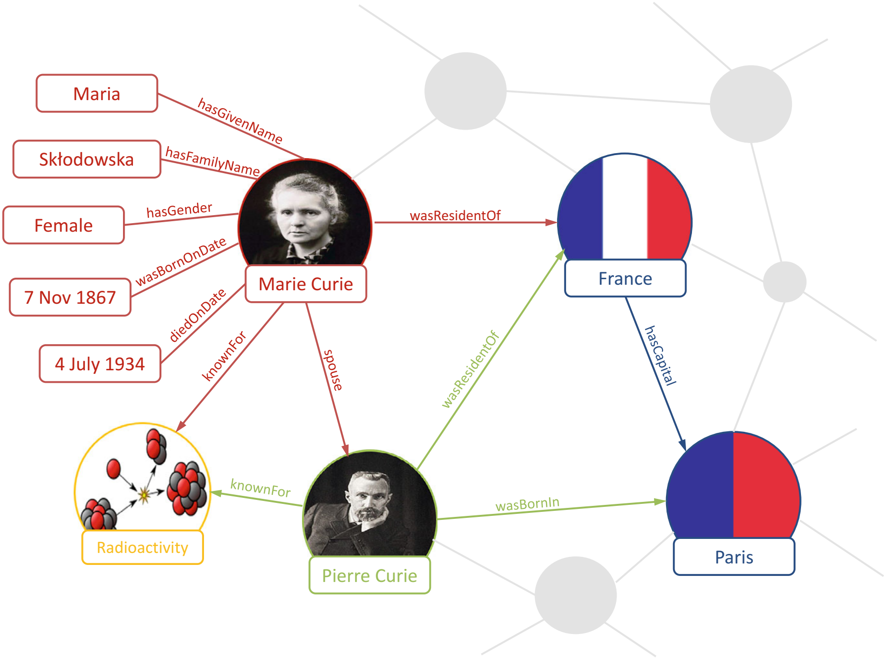


Fig. 2.6 Example of a knowledge graph for Marie Curie

Table 2.2 Facts of the knowledge graphs for Marie Curie

Subject	Predicate	Object
Marie Curie	hasGivenName	Maria
Marie Curie	hasFamilyName	Skłodowska
Marie Curie	hasGender	Female
Marie Curie	Spouse	Pierre Curie
Marie Curie	knownFor	Radioactivity
Marie Curie	wasBornOnDate	7 Nov 1867
Marie Curie	wasResidentOf	France
Marie Curie	diedOnDate	4 July 1934
France	hasCapital	Paris

Table 2.3 Size of some schema-based knowledge bases [45]

Knowledge graph	#of Entities	# of Relation types	# of Facts
Freebase	40 M	35,000	637 M
Wikidata	18 M	1632	66 M
DBpedia (en)	4.6 M	1367	538 M
YAGO2	9.8 M	114	447 M
Google Knowledge Graph	570 M	35,000	18,000 M

2.5.4 Smart Environment Example

An example of an entity-centric knowledge graph expressed as linked data within the context of a smart environment is illustrated in Fig. 2.7. Data and facts are specified as statements and are expressed as atomic constructs of a subject, predicate, and object, also known as a triple. The statement "Main Kitchen contains Coffee Machine" is expressed in the triple format as:

Subject—"Main Kitchen"
Predicate—"contains"
Object—"Coffee Machine"

RDF is designed for use in web-scale decentralised knowledge graph data models. For this reason, the statement parts need to be identified so that they can be readily and easily reused. RDF uses URIs for identification, so by expressing the previous statement in RDF it becomes:

http://data.deri.ie/rooms#r315
http://vocab.deri.ie/rooms#contains
http://water.deri.ie/devices#mr-coffee

URIs that describe the data can be uniformly used across systems, even if they come from different sources. The knowledge graph structure of the linked data, as illustrated in Fig. 2.7, easily supports optional parameters, and the evolution of parts of the data structure does not affect any other related data. The relations are described on a low-level; therefore, they combine (linking) pieces of data based on their relation types, and not only on their representation.

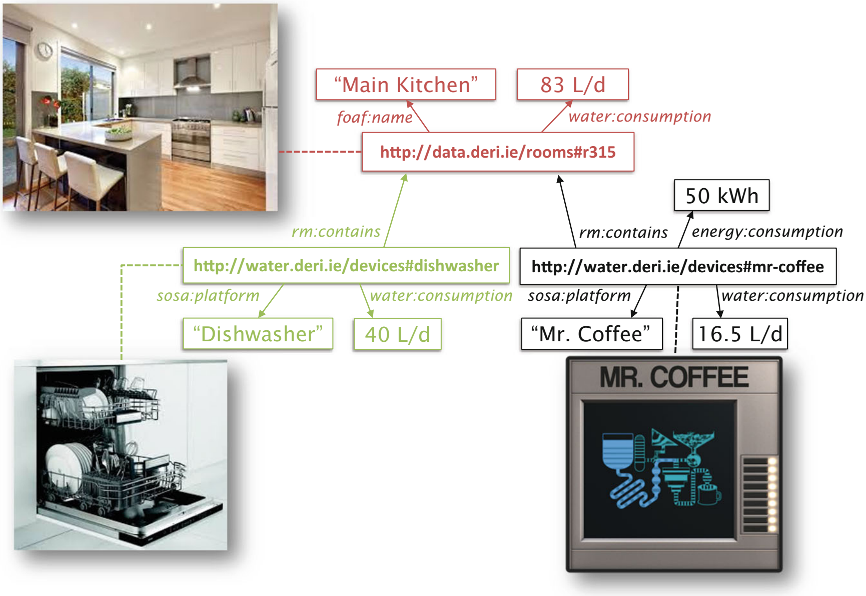


Fig. 2.7 Example of data linkage using URIs and RDF vocabularies in a smart environment

The flexibility to represent data and to support different relationships is a key benefit of linked data to support the sharing of data among systems. Linked data's use of vocabularies and ontologies is an important tool to establish shared meanings among different systems incrementally. This capability is critical to cross knowledge boundaries among systems with the use of knowledge graphs and entity-centric data integration to support the translation of knowledge.

2.6 Value: Continuous and Shared

The next part of the KVE Framework tackles value boundaries by identifying the common interests (data value chains) needed to support a value transformation for systems to share knowledge. We explore value disciplines and data network effects and how they can create new opportunities for the participants within the data ecosystem. These value opportunities can be the source of common interest to motivate the social, cultural, or business transformation needed to support knowledge exchange.

2.6.1 Value Disciplines

A value proposition is shaped by an underlying value discipline which describes different ways an organisation or system can differentiate itself from competitors. A strong value proposition can set the strategic focus that enables organisations or systems to set its vision and objectives. It can then tailor its value disciplines to match the need exactly. Treacy and Fred Wiersema [46] created a model to describe three generic value disciplines: (1) Operational Excellence, (2) Product Leadership, and (3) Customer Intimacy. The use of value disciplines has been explored in the broader areas of digital value [47], but also more specific areas such as open data [48]. Within the context of this work, we explore the use of data value disciplines to understand the value opportunities that are possible from data within a data ecosystem. The value of data within a smart environment can be considered in the context of the dynamics of knowledge-based organisations [23], where the processes of decision-making and organisational action are dependent on the process of sense-making and knowledge creation. Based on existing work [46–48], we identify the following three value disciplines for the participants (e.g. user, system, or organisation) of an intelligent systems data ecosystem:

- *Utility*: Tailors the value proposition to directly support the information needs of the participants. The objectives and information requirements of the participants should be defined to determine the usefulness of the data shared within the ecosystem. The utility can be shared between or can be unique to each participant.
- *Performance*: Tailors the value proposition to match to the needs of the participants specifically for improving processes for operational excellence. This can

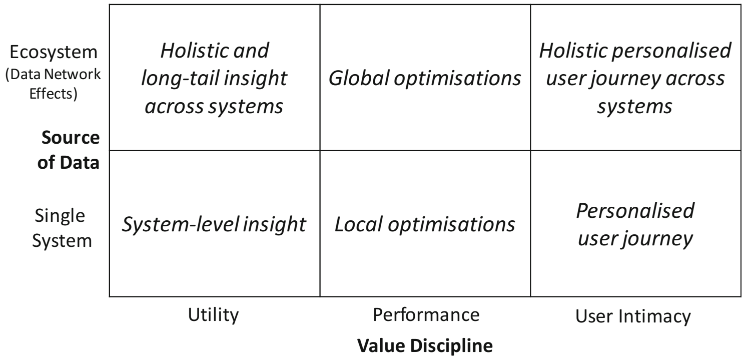
result in greater efficiencies with associated cost avoidance. Participants with this orientation aim to share data which support their primary performance objectives.

- *User Intimacy*: Tailors the value proposition to directly support the needs of users within the environment by providing information to enhance personalised user experiences and services.

2.6.2 Data Network Effects

A network effect is a positive effect described in economics and business where an additional user of a product or service has a positive effect on the value of that product/service to others. When a network effect is present, the value of the product or service increases according to the number of others using it. Robert Metcalfe popularised network effects (also called network externality or demand-side economies of scale) within the context of Ethernet as Metcalfe's law [49]. Within the area of data, network effects are starting to emerge, although in different forms, at both the data ecosystem and data product/service levels.

At the ecosystem level, the network effect can be seen as more systems/users join and contribute data to the data ecosystem; the overall data ecosystem becomes more valuable for the different value disciplines, see Fig. 2.8. Initiatives such as smart cities are showing how different sectors (e.g. energy and transport) can share data to maximise the potential for optimisation and value return. Data network effects occur at the data product/service level, where the data product/service becomes smarter (e.g. predictions, recommendations, and personalisation) as it gets more data from other participants. Leveraging data network effects requires a learning process within the data produce/ service that uses advanced analytics to extract insights from the collected data. The data network effect from cross-fertilisation of stakeholder and datasets from different sectors is a crucial element for advancing the big data economy in Europe [15] and is critical to support the value proposition of data ecosystems to their participants.



Source of Data	Utility	Performance	User Intimacy
Ecosystem (Data Network Effects)	*Holistic and long-tail insight across systems*	*Global optimisations*	*Holistic personalised user journey across systems*
Single System	*System-level insight*	*Local optimisations*	*Personalised user journey*

Fig. 2.8 Value transformation opportunities across value disciplines at the single system and ecosystem levels

2.7 Ecosystem: Governance and Collaboration

In order to understand some of the "political" and organisational issues that occur at ecosystem boundaries among systems, this section examines the work on business ecosystems to see how governance and collaboration can support knowledge flows. The section discusses data ecosystem coordination and the range of possible ecosystem design options.

2.7.1 From Ecology and Business to Data

The term Ecosystem was coined by Tansley in 1935 [50] to identify a basic ecological unit comprising of both the environment and the organisms that use it. Within the context of business, James F Moore [51–53] exploited the biological metaphor and used the term to describe the business environment. Moore defined a business ecosystem as an "economic community supported by a foundation of interacting organisations and individuals" [53]. A strategy involving a company attempting to succeed alone has proven to be limited regarding its capacity to create valuable products or services. It is crucial that businesses collaborate among themselves to survive within a business ecosystem [52, 54]. Innovation Ecosystems allow companies to create new value that no company could achieve by itself [55]; often, the ecosystem is centred around the technology platform or technology leadership of a focal firm.

The business ecosystem perspective is a more holistic way to look at the benefits of collaboration among companies, or in the case of a smart environment, the benefits of collaboration among systems. The ecosystem metaphor is, again, a useful metaphor to describe the data within and surrounding a smart environment. A data ecosystem is a socio-technical system enabling value to be extracted from data value chains supported by interacting organisations and individuals [15]. Data ecosystems can form in different ways around organisations, communities, technology platforms, or within or across sectors. Data ecosystems exist within many industrial sectors where vast amounts of data move among actors within complex information supply chains. Sectors with established or emerging data ecosystems include healthcare, finance [56], logistics, media, manufacturing, and pharmaceuticals.

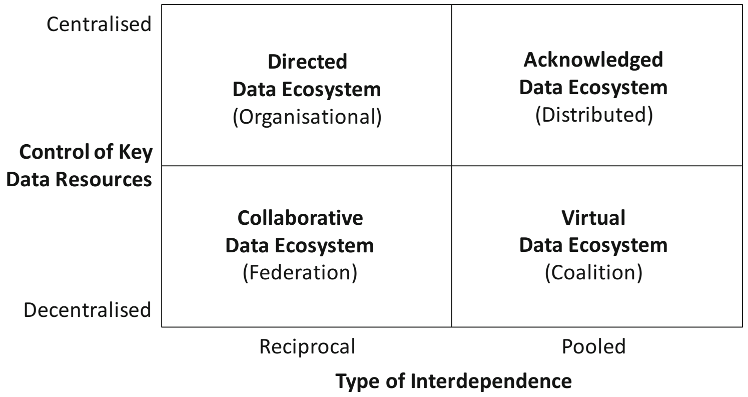
In natural ecosystems, smart organisms control their energy. In business ecosystems, a smart company manages information and its flows [57]. In data ecosystems, a smart company extracts the maximum value from the available data. The ecosystem can create the conditions for a marketplace competition among participants or enable collaboration among diverse, interconnected participants that depend on each other for their mutual benefit. Data ecosystems are useful for creating common interests among systems that are needed for the value transformation required to share data. The benefits of sharing and linking data across domains and industry sectors are becoming apparent with the potential for new value opportunities on the Web of Data.

2.7.2 The Web of Data: A Global Data Ecosystem

The web is moving from a medium for sharing documents to a medium that can also be used to share data. Fuelled by the *Open Data* initiative, the emerging “Web of Data” means easier access to data for users. Typically, open data is non-textual material such as maps, genomes, chemical compounds, mathematical, and scientific formulae. Open data can also include generalised business news, product information, and financial data [56] available from an assortment of sources. Demands for higher levels of transparency have resulted in Open Government initiatives that have made available large numbers of statistical, financial, and economic datasets for public consumption. A number of large-scale knowledge bases have been made available from both private and not-for-profit initiatives, including Google Knowledge Graph, DBpedia, and YAGO, to name a few. The LinkedIn Economic Graph describes all the data on LinkedIn like companies, members, and jobs, to provide a digital representation of the global economic activity with a focus on employment opportunities. The Linked Open Data Cloud represents a large number of interlinked RDF datasets within the broader ecosystem that is being actively used by industry, government, and scientific communities [58]. The linked data cloud has been growing in the past years and provides a foundation upon which applications can be built. The Facebook Open Graph describes a rich object in a social graph, simplifying the process of sharing social data on the web. The Schema.org imitative was founded by Google, Microsoft, Yahoo, and Yandex to create shared vocabularies through an open community process for publishing data. Schema.org vocabularies can be used with many different encodings, including RDFa, Microdata, and JSON-LD. These vocabularies cover entities, relations among entities and actions, and can easily be extended through a well-documented extension model. Each of these initiatives is part of a broader data ecosystem in the emerging Web of Data.

2.7.3 Ecosystem Coordination

Within the KVE Framework, the role of the data ecosystem and data value chains is to support the value transformations necessary (social, cultural, value) to create new common interests and value opportunities among intelligent systems. To achieve this, it is necessary to understand and support the social, political, and organisational changes needed for coordinating ecosystems. To understand the dynamics of an intelligent systems data ecosystem we can look into the literature on System of Systems [28] and Business Ecosystems [59] to enable us to understand the data ecosystems that can exist [1]. In Fig. 2.9, we can see the different types of data ecosystems that can form around intelligent systems within a smart environment. Two critical criteria that influence the design of a data ecosystem and the relationships among participants are:



Control of Key Data Resources	Reciprocal	Pooled
Centralised	**Directed Data Ecosystem** (Organisational)	**Acknowledged Data Ecosystem** (Distributed)
Decentralised	**Collaborative Data Ecosystem** (Federation)	**Virtual Data Ecosystem** (Coalition)

Type of Interdependence

Fig. 2.9 Topology of data ecosystems [1]. Adapted from [59, 28]

- *Control of Key Data Resources*: Who controls the essential data resources in the data ecosystem? Does a single "Keystone" [57] actor control the key data resources that all others depend on? Alternatively, is control of the critical data resources spread across multiple actors in the data ecosystem?
- *Participant Interdependence*: The degree to which different participants in the data ecosystem must interact and exchange data for performing their activities. Reciprocal interdependence requires high levels of coordination among the participants, while pooled interdependence enables loose coupling among participants.

Drawing inspiration from the SoS classification by Maier [28] (including Virtual, Collaborative, Acknowledged, and Directed) and the business ecosystem topology by Koenig, we can (see Fig. 2.9) consider the different types of data ecosystems [1] that may exist within a smart environment and the nature of the relationships among the participants.

- *Directed Data Ecosystems*: Centrally controlled to fulfil a specific purpose. Typically found within an organisation setting or following a keystone model. Participants within a directed data ecosystem maintain an ability to operate independently, but their normal operational mode is subordinated to the centrally managed purpose of the data ecosystem.
- *Acknowledged Data Ecosystems*: Have defined objectives and pooled dedicated resources. The constituent systems retain their independent ownership and objectives. Changes in the data ecosystem are based on collaboration among the distributed participants.
- *Collaborative Data Ecosystems*: Participants interact voluntarily to fulfil an agreed-upon central purpose. The primary players collectively decide the means of enforcing and maintaining standards among the federations of participants.

- *Virtual Data Ecosystems*: Have no central management authority and no centrally agreed-upon purpose. Bottom-up coalitions of participants emerge from a virtual data ecosystem to pool decentralised resources to achieve specific goals.

Within a well-functioning data ecosystem, the participants are efficiently and effectively sharing across knowledge, value, and ecosystem boundaries. The nature of the ecosystem and the systems and their dynamics will affect the design and operation of the data ecosystem. To enable an intelligent systems data ecosystem, it is clear we will need to rethink some of the fundamentals of current intelligent system design approaches regarding governance, economics, and technical approaches.

2.7.4 Data Ecosystem Design

Several ecosystem design characteristics are detailed in Table 2.4. It is worth considering that multiple data ecosystems could exist at one time, and the operation of a data ecosystem can change depending on the circumstances. Concerning the design of intelligent systems, these design characteristics can affect the style of infrastructure that is needed to support data sharing within the data ecosystem, from data provided by a single dominant actor on their proprietary infrastructure, to a community, pooling their data in a managed open source data platform.

Table 2.4 Data ecosystem design space

Design characteristics		Solution design space	
Governance	Control	Centralised	Decentralised
	Interdependence	Reciprocal	Pooled
	Structure	Authoritarian	Democratic
	Regulation	None	Enforceable
	Independence	Controlled	Autonomous
	Environment	Stable	Dynamic
Economic	Model	Pay	Free/sharing
	Connectivity	Keystone	Value network
	Data market	Single-sided	Multi-sided
	Collaboration	Competition	Cooperation
Technical	Infrastructure	Proprietary	Open
	Data availability	Closed	Open
	Privacy	Monitoring	Privacy-protecting
	Data formats	Homogeneous	Heterogeneous
	Data services	Exact	Approximate

2.8 Iterative Boundary Crossing Process: Pay-As-You-Go

The *Iterative Boundary Crossing Processes* follow a socio-technical approach to accommodate iterations at crossing system boundaries. A key capability is the need to support a flexible, iterative approach that facilitates incremental agreements and investments among stakeholders. Pay-as-you-go data management approaches (such as dataspaces) are needed for technical concerns, while data ecosystem supports are needed to facilitate incremental transformations of political and organisational concerns.

2.8.1 Dataspace Incremental Data Management

A dataspace is an emerging approach to data management that is distinct from current approaches. The dataspace approach recognises that in large-scale integration scenarios, involving thousands of data sources, it is difficult and expensive to obtain an upfront unifying schema across all sources [2]. Within the dataspace paradigm, data management pushes the boundaries of traditional databases in two main dimensions [2]: (1) Administrative Proximity, which describes how data sources within a space of interest are close or far in terms of control; and (2) Semantic Integration, which refers to the degree to which the data schemas within the data management system are matched up. Dataspaces shift the emphasis to providing support for the co-existence of heterogeneous data that does not require a significant upfront investment into a unifying schema. Data is integrated on an "as-needed" basis with the labour-intensive aspects of data integration postponed until they are required. Dataspaces reduce the initial effort required to set up data integration by relying on automatic matching and mapping generation techniques. This results in a loosely integrated set of data sources. When tighter semantic integration is required, it can be achieved in an incremental "pay-as-you-go" fashion by detailed mappings among the required data sources. Dataspaces are described in further detail in Chap. 3. We have created the Real-time Linked Dataspace (RLD) (see Chap. 4) as a data platform for intelligent systems within smart environments. The RLD combines the pay-as-you-go paradigm of dataspaces with linked data and real-time stream and event processing capabilities to support a large-scale distributed heterogeneous collection of streams, events, and data sources [4].

The KVE Framework has outlined a high-level approach to support the exchange of knowledge among intelligent systems within a data ecosystem. In order to realise the sharing of knowledge between interconnected intelligent systems, there is a need for a data platform.

2.9 Data Platforms for Intelligent Systems Within IoT-Based Smart Environment

Platform approaches have proved successful in many areas of technology [60], from supporting transactions among buyers and sellers in marketplaces (e.g. Amazon), innovation platforms which provide a foundation on which to develop complementary products or services (e.g. Windows), to integrated platforms which are a combined transaction and innovation platform (e.g. Android and the Play Store).

The idea of large-scale "data" platforms have been touted as a possible next step for the development of smart environments [1] and data ecosystems. An ecosystem data platform would have to support continuous, coordinated data flows, seamlessly moving data among intelligent systems. The design of infrastructure to support data sharing and reuse is still an active area of research. In order to understand the general requirements necessary to share data, we examine the "FAIR Data" principles [61] that have been defined to support data reuse within the scientific community. Then, to understand the specific data sharing requirements for an intelligent systems data ecosystem, we examine the data management needs of five different IoT-based smart environments.

2.9.1 FAIR Data Principles

In order to improve the data infrastructure supporting the reuse of research data, a group of stakeholders from academia, industry, funding agencies, and research publishers proposed a set of principles known as the FAIR Data Principles [61]. The FAIR principles are Findability, Accessibility, Interoperability, and Reusability with a detailed breakdown of the principles provided in Table 2.5. The objective of the principles is to act as a set of guidelines to data producers and publishers to maximise the reusability of research data. The FAIR principles are designed to enable proper data management to support knowledge discovery and innovation, and the subsequent data and knowledge integration and reuse. The principles define the goals of good data management and stewardship practices to improve its reusability. The principles can influence the design of algorithms, tools, and workflows for research data. The broad application of the principles can lead to a data research ecosystem that supports extracting maximum benefit from research investments by ensuring transparency, reproducibility, and reusability. Within the context of this work, we use the principles as a high-level guide for the design of a data platform to support knowledge sharing between intelligent systems within a smart environment.

Table 2.5 FAIR guiding principles for scientific data management and stewardship [61]

To be findable
F1. (Meta)data is assigned a globally unique and persistent identifier
F2. Data is described with rich metadata (defined by R1 below)
F3. Metadata clearly and explicitly includes the identifier of the data it describes
F4. (Meta)data is registered or indexed in a searchable resource
To be accessible
A1. (Meta)data is retrievable by their identifier using a standardised communications protocol
A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
A2. Metadata is accessible, even when the data is no longer available
To be interoperable
I1. (Meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (Meta)data uses vocabularies that follow FAIR principles
I3. (Meta)data includes qualified references to other (meta)data
To be reusable
R1. (Meta)data is richly described with a plurality of accurate and relevant attributes
R1.1. (Meta)data is released with a clear and accessible data usage license
R1.2. (Meta)data is associated with detailed provenance
R1.3. (Meta)data meets domain-relevant community standards

2.9.2 Requirements Analysis

Over the past years, we have been involved in a number of projects [4, 18, 62, 63] concerned with next-generation data platforms for intelligent systems within smart environments. The smart environments focused on intelligent energy and water management with varying sizes of data ecosystems. The five pilots are:

- *Smart Airport*: Linate airport in Milan represents large-scale commercial energy and water consumer for use from washing activities, toilets, restaurants, and irrigation, flight operations, to safety-critical infrastructure for emergency response. Linate targets a variety of users, from the company's employees (including executives, operational managers, and technical staff), to passengers. The variety of sensors used in the airport requires the management of heterogeneous events and their availability to applications in near-real-time. Significant contextual data from the airport's operational legacy systems is needed to process the events for decision-making.
- *Smart Office*: The Insight Building was built in the 1990s without a building management system and has been retrofitted with energy sensors. As typically in an organisation, Insight has several information systems that run its operations, including finance and enterprise resource planning, budgeting, and office IT assets. These enterprise systems can help in identifying energy wastage and promoting conservation actions within the office.

- *Smart Homes*: The Municipality of Thermi in Greece provides a residential smart water pilot with a representative sample of ten domestic residences. The target users are the residents (both adults and children), municipality management, a developer community for smart home "Apps," research scientists, and the local water utility. Data from IoT devices in each home needs to be managed in a near-*real-time* manner to provide feedback to users on their water consumption. Secure sharing of data with both the research and developer community is needed.
- *Mixed Use*: The Engineering Building at NUI Galway in Ireland is a state-of-the-art smart building with significant numbers of sensors and actuators. Target users include academic staff, managers, technicians, researchers, and students. This smart environment is designed to be a "living laboratory" where the building itself is an interactive teaching tool where students can utilise data from the environment in their projects and research works. Making data easily reusable by occupants in the environment is an essential requirement.
- *Smart School*: Coláiste na Coiribe is a newly constructed Irish language secondary school. The school accommodates students aged between 12 and 18, together with teaching and operational staff. The school has been fitted with a commercial state-of-the-art building management system to manage its energy and water consumption. A key challenge is to customise the communication of energy and water data for the diverse range of school stakeholders.

For each of these five smart environments, a system analysis was performed to identify the functional and non-functional information processing and sharing requirements. These requirements complement the FAIR principles by including concrete requirements for data processing, querying, and data ecosystem support, including the need for iterative, incremental processes. The following common data platform requirements were identified across the pilots [4]:

- *Pay-As-You-Go Data Integration, Accessibility, and Sharing*: Each smart environment contains potentially thousands of data sources from sensors and things to legacy information systems. Harnessing this data is critical to enabling the smart environment. Challenges include the integration of multiple formats and semantics, discoverability and access, and data re-use and sharing in a low-cost and incremental manner [33, 64–67]. This high-level requirement can be broken down into a set of technical requirements:
  - *Standard data syntax, semantics, and linkage:* Facilitate integration and sharing, ideally with open standards and non-proprietary approaches.
  - *Single-point data discoverability and accessibility:* Allow the organisation and access to datasets and metadata through a single location.
  - *Incremental data management:* Enable a low barrier to entry and a pay-as-you-go paradigm to minimise costs.
- *Secure Access Control*: Support data access rights to preserve the security of data and privacy of users in the smart environment. Access control is needed at both the level of the data source and at the level of the data item (i.e. entity-value).

- *Real-time Data Processing and Historical Querying*: Each environment requires support for the real-time processing of data generated from sensors and things within the environments. This requirement can be broken down into two technical requirements:
  – *Real-time data processing:* Including ingestion, aggregation, and pattern detection within event streams originating from sensors and things in the smart environment.
  – *Unified querying of real-time data and historical data:* Provide applications and end-users with a holistic queryable state of the smart environment at a latency suitable for user interaction.
- *Entity-Centric Data Views*: Intelligent applications and end-users need to be able to explore and query the data from an entity perspective, such as energy or water usage in a specific building zone. The raw data generated by things (e.g. a smart tap) within the environments often only report on the observed values of a property (e.g. water consumption). Thus, the raw sensor/thing data may require additional contextual information, such as the location of the sensor [64–66]. This high-level requirement can be broken down into two technical requirements:
  – *Entity management:* The storage, linkage, curation, and retrieval of entity data, such as users, zones, and locations
  – *Event enrichment:* Enhancement of sensor/things streams with contextual data (e.g. entities) to make the stream data more encapsulated and useful in downstream processing

The level of importance of these common data requirements varies within each pilot as detailed in Table 2.6. Many other requirements were identified within the smart environments, including interoperability of devices and network protocols, user profiling, the resilience of remote sensors, and advanced privacy-preserving analytics.

Table 2.6 Level of importance of common data platform requirements [4]

Requirements	Smart Airport	Smart Office	Smart Home	Mixed Use	Smart School
Standard data syntax, semantics, and linkage	High	Medium	Low	Medium	Medium
Single-point data discoverability and accessibility	High	Medium	High	High	Medium
Incremental data management	High	High	Low	High	Medium
Secure access control	High	High	High	High	Medium
Real-time data processing	High	High	Medium	High	High
Unified querying of real-time data and historical data	High	High	High	High	High
Entity management	High	High	Medium	High	Medium
Event enrichment	High	High	High	High	Medium

2.10 Summary

The digital transformation is creating a data ecosystem with data on every aspect of our world spread across a range of information systems. Data ecosystems present new challenges to the design of intelligent systems and System of Systems that demands a reconsideration of how we deal with the needs of large-scale, data-rich smart environments. In this chapter, we have explored the barriers to the sharing of knowledge among intelligent systems within a smart environment and how they can be overcome with the capabilities within the Knowledge Value Ecosystem (KVE) Framework. The implementation of these capabilities was explored using linked data, knowledge graphs, and data value chains, which provide solid foundations for tackling system boundaries of knowledge exchange among systems. Finally, the chapter examined the need for data platforms to support the sharing of data between intelligent systems within a data ecosystem and identified common data platform requirements.