Chapter 18
Future Research Directions for Dataspaces, Data Ecosystems, and Intelligent Systems

Keywords Dataspaces · Data ecosystems · Intelligent systems · Research challenges · Technology adoption · Trusted data sharing · Governance · Incremental systems engineering · Human-centricity

18.1 Introduction

As we move toward 2030, today's computing paradigms such as data-intensive computing (Big Data), Open Data [380], Knowledge Graphs, Machine Learning, Large-Scale Distributed Systems [381], Internet of Things (IoT), Physical-Cyber-Social Computing [14], Service-Oriented [382], and Cloud/Edge Computing [383] will be the foundations to the realisation of the vision of intelligent systems. In fact, real-world intelligent systems are being enabled by a combination of these paradigms using a mixture of architectures (centralised, decentralised, and a combination of both) and infrastructures such as Middleware and IoT platforms to support the development of intelligent systems and applications [13, 67, 295, 384].

This chapter begins in Sect. 18.2 with an examination of what is required for the widespread adoption of the dataspace approach. Next, in Sect. 18.3 the chapter explores the research landscape towards 2030 by identifying the principal research directions for the dataspaces, data ecosystems, and intelligent systems, including large-scale decentralised support services, multimedia/knowledge-intensive event processing, trusted data sharing, data governance, and economic models, evolving intelligent systems engineering and cognitive adaptability, and finally the path towards human-centric systems. The chapter finishes with a summary in Sect. 18.4.

E. Curry, *Real-time Linked Dataspaces*,
https://doi.org/10.1007/978-3-030-29665-0_18

18.2 Dataspaces: From Proof-of-Concept to Widespread Adoption

Our vision for the intelligent systems of 2030 is that they will be a significant part of a dynamic global data ecosystem where vast amounts of data can move among actors within complex information supply chains [1, 25]. Users, applications, and machines will still need to leverage these data flows to optimise physical and virtual systems in the areas of economy, environment, energy, water, waste, people (intellectual endowment and engagement), built environment, mobility (transportation), and public spaces [385, 386]. We believe dataspaces will be a crucial technology platform, enabling users to harness the data from the global data ecosystem. In order to deliver the potential of dataspaces, the wide-scale diffusion of the technology is necessary. The literature on the diffusion of innovations and technology adoption [387] can assist in understanding how this could happen.

In their study of technology change, Anderson and Tushman [388] argue that technology progresses in a series of cycles, hinging on technological discontinuity followed by a design competition which results in the emergence of a dominant design. According to Anderson and Tushman, the dominant design is never in the same form as the first discontinuity, and it is not on the leading edge of technology; it bundles features to meet the requirements of most of the market. Once a dominant design emerges, organisations often cease to invest in learning alternative designs and instead focus on developing competencies related to the dominant design.

Understanding these cycles and patterns can indicate as to the trends in the data management domain and the potential to improve the adoption of the dataspace paradigm. The first wave of dataspace initiatives [2, 87, 179] can be seen as a large-scale design competition consisting of Proof-of-Concept projects that explored the potential for dataspaces within specific data management use cases. The goal was to understand the requirements, explore the design space, and discover the boundaries of the many different support services needed to enable the dataspace data management paradigm. The early dataspaces were designed and developed by world-leading researchers and required high levels of expertise. The defining characteristics of many projects in this wave was a focus on the experimental design together with a pilot deployment (e.g. biomedical, energy [100], personal [87, 88, 91, 92]) to meet specific data management requirements.

The second wave of dataspace initiatives, now underway [4, 101], is focusing more on general deployments of dataspaces beyond the specific initial use cases to drive broader adoption. The key challenge in this wave is the need to identify the dominant design needed to support the requirements of mass-market adoption. The innovation adoption literature can again guide dataspace researchers in improving the uptake of their technology. The likelihood of an innovation being adopted can be increased if it possesses specific key characteristics [389]. The following criteria have been adapted from [389] for the context of a dataspace within a data ecosystem:

- *Relative Advantage:* Enabling a better functioning data ecosystem and usage of data within the ecosystem.
- *Compatibility:* The degree to which a dataspace is consistent with existing stakeholder values, or interests, and usage context.
- *Complexity:* The degree of difficulty involved in implementing the dataspace and communicating benefits to stakeholders.
- *Trialability:* The degree to which experimentation is possible with a dataspace.
- *Cost Efficiency and Feasibility:* Concerning existing comparable data management practices.
- *Evidence:* The availability of research evidence and practical efficacy of a dataspace.
- *Risk:* Level of risk associated with the implementation and adoption of a dataspace.

Dataspaces will need to possess a number of these characteristics if they are to be successfully adopted within the general data management domain. This sets out a clear research direction for next-generation dataspaces.

18.3 Research Directions

Dataspaces are a relatively new research area that brings together several other areas in computer science and other disciplines. We now discuss research areas [1, 4] which are essential to enabling the next-generation of dataspaces, data ecosystems, and intelligent systems (see Fig. 18.1). Research is needed to overcome many challenges, including decentralised support services, support for multimedia data, trusted data sharing, governance and economic models, incremental systems engineering, and human-centricity.

18.3.1 Large-Scale Decentralised Support Services

As dataspaces are deployed at larger scales, it will be necessary to create enhanced support services, to scale entity management, and to minimise the cost of operation for these deployments [4]. Challenges include:

- *Enhanced Supported Services:* Many enhancements are possible for support services of dataspaces including the use of natural language interfaces to improve user experience [390], decentralised support services for large-scale deployments [391], and privacy-by-design [392] approaches to support the ever-increasing amount of personal information captured in intelligent systems.
- *Scaling Entity Management:* Within larger-scale deployments, it will be necessary to enhance the entity management services to support both the increase in entities, data, and users. Ranking and summarisation need to be query- and

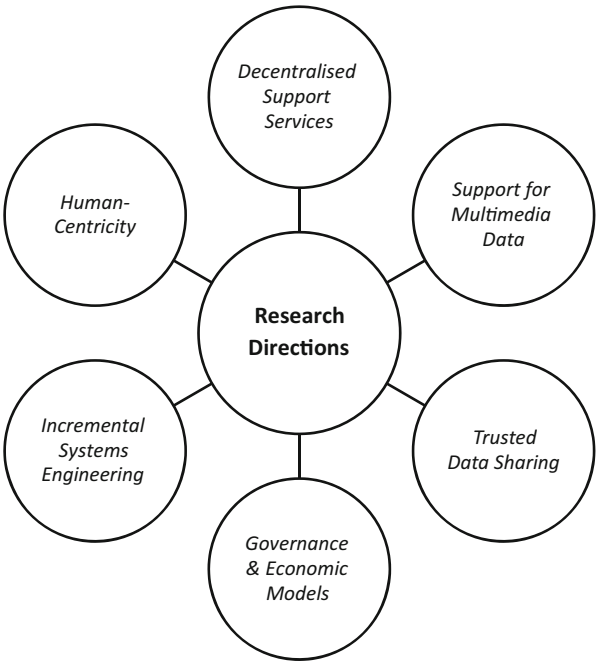

Fig. 18.1 Research directions for dataspaces, data ecosystems, and intelligent systems

activity-relevant [86], with relevant facts, but at the same time diverse to cater for a wide range of information/conceptualisation need. A trade-off between processing time and expressiveness is necessary. Furthermore, there is significant potential for extensive usage of large-scale crowdsourcing for "human-in-the-loop" data management and curation [71, 117].

- *Maintenance and Operation Cost:* As the size of deployment increases, it will be necessary to investigate new techniques to improve the performance of dataspaces in terms of the maintenance and operational costs of the support platform within large-scale deployments (e.g. city-level) [26].

18.3.2 Multimedia/Knowledge-Intensive Event Processing

As multimedia streams become more pervasive with the Internet of Multimedia Things (IoMT), it will be necessary for dataspaces to provide specific support services to process and manage these streams. New techniques and approaches will be needed for:

- *Support Services for Multimedia Data:* As multimedia data becomes more common in data ecosystems through the IoMT, there will be a direct need for appropriate support services within dataspaces. There is an opportunity to

leverage advances in deep learning for image processing (e.g. object detection) [30], which can be the basis of dataspace support services for rich content types including text and multimedia streams.

- *Placement of Multimedia Data and Workloads:* The increased computing resources needed to process, and extract multimedia data will pose challenges for existing techniques for processing and data placement. This will require dataspace support services to consider the simultaneous training and processing of multimedia streams, taking into consideration the geospatial and temporal characteristics of smart environments.
- *Adaptive Training of Classifiers:* To effectively process multimedia data, dataspaces will need to be able to assemble the appropriate classifiers to extract features from multimedia content based on the needs of users of the dataspace at runtime. The training of classifiers needs to be adaptable to the changing requirements of data ecosystems. There is a need to support transfer learning among intelligent systems and for collective efforts to build pre-trained models for datasets and to bootstrap dataspace support services. Finally, distributed approaches to training classifiers are needed to maximise the available resources from the cloud to the edge of the network.
- *Complex Multimedia Event Processing:* To detect patterns from multimedia streams within a dataspace, it will be necessary to investigate new techniques for complex multimedia event processing. Challenges include defining the language to express the complexity of the event patterns and the content of the event, optimisation techniques to improve system performance for event detection over computationally intensive multimedia streams, and methods to train models over incoming media streams for new unseen queries in lack of available training data.

18.3.3 Trusted Data Sharing

There is a need to enable the trusted sharing of data among organisations, people, and systems. This will pose significant challenges for:

- *Trusted Platforms:* A trusted data platform focuses on the secure data sharing among a group of participants (e.g. industrial consortiums sharing private or commercially sensitive data) within a clear legal framework. An ecosystem data platform would have to be infrastructure agnostic and must support continuous, coordinated data flows, seamlessly moving data among systems [1]. Data exchange could be based on models for monetisation or reciprocity. Data platforms can create possibilities for smaller organisations and even individual developers to get access to large volumes of data, enabling them to explore their potential. Trusted platforms open many research areas for dataspaces, including data discovery, curation, linking, synchronisation, standardisation, and decentralisation [25].

- *Usage Control:* The challenges with data sharing go beyond technical issues to issues of data ownership, privacy, business models, smart contracts, and licensing and authorised reuse by third parties. The control paradigm for shared data must shift from today's *access* control to *usage* control, and dataspaces will need to support both of these usage control for both organisations and individuals.
- *Personal Dataspaces:* There is a need for personal dataspaces for the management of the data of the individual. Personal dataspaces will need to respect the relevant legislation for personal data (e.g. General Data Protection Regulation) and allow an individual to remain in control of their personal data and its use. Personal dataspaces will need to balance the need for privacy with the benefits of analytics and handle this trade-off based on the preferences of the individual. Techniques for preserving privacy for metadata, query privacy, and privacy-preserving integration of independent data sources will all be needed in next-generation dataspaces.
- *Industrial Dataspaces:* The sharing of data among commercial organisations will also increase. Industrial dataspaces [101] will be needed to facilitate the trusted and secure sharing and trading of commercial data among collaborating organisations. These platforms will need to provide support services that enable a data marketplace that facilitates the automated licensing of data exchanged among organisations and the enforcement of legal rights and appropriation of remuneration to the original data owners.

18.3.4 Ecosystem Governance and Economic Models

For mass collaboration to take place within data ecosystems, we need to overcome the challenges of dealing with large-scale agreements among potentially decoupled interacting parties [1]. New approaches will be needed for:

- *Decentralised Data Governance:* Research is needed on decentralised data governance models for data ecosystems that support collaboration and fully consider ethical, legal, and privacy concerns. Data governance within a data ecosystem must recognise data ownership, sovereignty, and regulation while supporting economic models for the sustainability of the data ecosystem [1]. A range of decentralised governance approaches may guide a data ecosystem from authoritarian to democratic, including majority voting, reputation models (e.g. eBay), proxy-voting, and dynamic governance (e.g. sociocracy: circles and double linking) [393]. Dataspaces will need to enforce these data governance models automatically.
- *Economic Models:* Economic model may be used as an incentivisation factor within governance models including support for "data-vote exchange" models (pay for votes with data), and economic models for peer-to-peer systems [394, 395]. The sharing and exchange of data within dataspaces could also be based on models for monetisation or reciprocity. Data platforms can create

possibilities for smaller organisations and even individual developers to get access to large volumes of data, enabling them to explore their potential [1].

18.3.5 Incremental Intelligent Systems Engineering: Cognitive Adaptability

The design of adaptive intelligent systems will need to consider the implication of operating within a smart environment and its associated data ecosystem [1]. This will pose significant challenges for systems engineering:

- *Pay-As-You-Go Systems:* The boundaries of systems will be fluid and will change and evolve at run-time to adapt to the context of the current situation. However, we must consider the cost of system participation, and support "pay-as-you-go" approaches at both the system and data levels [1]. How can the pay-as-you-go approach of dataspaces be extended to the design of incremental and evolving systems? How can we integrate systems on an "as-needed" basis with the labour-intensive aspects of system integration postponed until they are required?
- *Cognitive Adaptability:* Work on evolving systems engineering [29] will need to consider the inclusion of data-driven probabilistic techniques that can provide "Cognitive Adaptability" to help intelligent systems adapt to changes in the environment that were unknown at design-time by enriching the control-loop with observational data from the environment. Intelligent system designers will need to consider the varying levels of accuracy offered by data-driven approaches, providing best-effort or approximate results using the data accessible at the time [1]. How can we mix deterministic and statistical approaches in the design of intelligent systems? How can we test and verify these systems? There is a need to support transfer learning among intelligent systems and for joint efforts to build pre-trained models for system adaptability. Dataspaces can play a role in supporting these collective efforts.

18.3.6 Towards Human-Centric Systems

Currently, intelligent systems make critical decisions in highly engineered systems (e.g. autopilots) where users receive specialised training to interact with them (e.g. pilots). As we move forward, intelligent systems will be making both critical and lifestyle decisions: from the course of treatment for a critical illness, safely driving a car, to choosing what takeout to order and the temperature of our shower [1]. This will pose specific challenges in the design of human-centric systems:

- *Explainable Artificial Intelligence and Data Provenance:* Data-driven decision approaches (including Cognitive and AI-based techniques) will need to provide

explanations and evidence to support their decisions and guarantees for the decisions they recommend. How can we trust the large-scale, data-driven decision-making provided by dataspace-powered AI platforms? This will require a greater need for provenance support within dataspaces to support the audit trail necessary to justify a data-driven decision [396].

- *Human-in-the-Loop:* The role of users in intelligent systems will not be a passive one. Users are a critical part of socio-technical systems, and we need to consider more ways of including the "Human in the Loop" within future intelligent systems. Active participation of users can improve their engagement and sense of ownership of the system. Indeed, active involvement of the user could be a condition for them granting access and usage of their private data. Research is needed to give trust in algorithms and data, in the trusted co-evolution between humans and AI-based systems, and in the legal, ethical, and privacy issues associated with making data-driven critical decisions [1].

18.4 Summary

This chapter examined what is required for the widespread adoption of the dataspace approach. The chapter then explored the future research landscape by identifying the principal research directions for the dataspaces, data ecosystems, and intelligent systems including large-scale decentralised support services, multimedia/knowledge-intensive event processing, trusted data sharing, data governance and economic models, evolving systems engineering and cognitive adaptability, and finally the path towards human-centric systems.