Journal of Smart Cities and Society 1 (2022) 31–51
DOI 10.3233/SCS-210119
IOS Press

Data governance in smart cities: Challenges and solution directions

Sunil Choenni [a,b,*], Mortaza S. Bargh [a,b], Tony Busker [b] and Niels Netten [a,b]

[a] *Creating 010, Rotterdam University of Applied Sciences, Rotterdam, The Netherlands*
[b] *Research and Documentation Centre (WODC), Dutch Ministry of Justice and Security, The Hague, The Netherlands*

Received 28 July 2021
Accepted 11 January 2022

Abstract. Today, our environment and the objects therein are equipped with an increasing number of devices such as cameras, sensors, and actuators, which all together produce a huge amount of data. Furthermore, we observe that citizens generate data via social media applications running on their personal devices. Smart cities and societies are seeking for ways to exploit these vast amounts of data. In this paper, we argue that to take full advantage of these data, it is necessary to set up data governance properly, which includes defining, assigning, and allocating responsibilities. A proper setting up of data governance appears to be a challenging task since the data may be used irresponsibly, thoughtlessly and maliciously, resulting in many (un)wanted side effects such as violation of rules and regulations, human rights, ethical principles as well as privacy and security requirements. We elaborate on the key functionalities that should be included in the governance of a data ecosystem within smart cites, namely provisioning the required data quality and establishing trust, as well as a few organizational aspects that are necessary to support such a data governance. Realizing these data governance functionalities, among others, asks for making trade-offs among contending values. We provide a few solution directions for realizing these data governance functionalities and making trade-offs among them.

Keywords: Data governance, data eco system, data quality issues, trust issues, fair principle

1. Introduction

The fields of (big) data science and artificial intelligence are undergoing enormous developments, which result into an enormous growth of data driven applications. These developments and applications are crucial for shaping the concept of smart cities, which aims at providing a sustainable, livable, and inclusive environment for their residents, businesses, and visitors. To achieve this, a smart city, in a sense, utilizes (big) data applications to interconnect and integrate information generated by digital infrastructures embedded in cities' physical environment. However, this comes with several challenges, such as dealing with the risks of privacy and security breaches, misguidances,

*Corresponding author: Sunil Choenni, Creating 010, Rotterdam University of Applied Sciences, Wijnhaven 107, 3011WN Rotterdam, The Netherlands. E-mail: sunilchoenni@gmail.com.

and misinterpretations. To fully benefit from the potential of these applications care should be taken for the entailed challenges. Since data are a cornerstone for these applications, the above mentioned challenges can be linked to data related challenges and, in turn, should be addressed adequately. These data related challenges range from data quality, security and privacy issues to the semantics and interpretation of data [16,18].

These challenges are not only caused by the "secondary use" of the data, i.e., the data are used for another purpose than the originally collected for one, but also by the fact that data should be shared with others (i.e., used primarily or secondarily) in an appropriate way (i.e., with good enough quality) to serve the purpose in mind and fully exploit the potentials of the data for that purpose. For example, the primary goal of checking in and checking out in the public transportation system is to bill the passenger for the traveled distance. However, the (aggregated) set of these data can also be used for a better planning of public transportation. If it appears that at specific hours of a day many people are checking in on a route, then longer vehicles might be deployed on this route at these hours. The same data can be shared with the police in order to optimize their police patrols by identifying the hotspots in a city. The data might also be exploited by commercial companies to develop new services for citizens, such as (real-time) route planners to guide citizens along the safest and shortest way to their destination. For municipalities these data may be the basis to develop their environmental policies, e.g., planting trees against air pollution. However, data sharing raises several questions, such as to whom and on at what aggregation level the data can or should be shared, whether the data quality is good enough for the secondary usage in mind, how to prevent, for example, function creep, privacy and security breaches whilst sharing the data.

Given, on the one hand, the importance of data sharing in a smart city and, on the other hand, the increased complexity involved with data sharing among (many) stakeholders, we argue the need for establishing appropriate data ecosystems. Collaboration and partnerships between several stakeholders are needed to face the increased complexity in data sharing to meet the goals of a smart city. In such a data ecosystem, all stakeholders are involved to shape the data governance within the whole system, i.e., defining, assigning, and allocating the responsibilities according to the norms, values and objectives of smart cities. These responsibilities include e.g., guaranteeing the quality of data, securing the storage and exchange of data, optimizing the tradeoff between the data utility and data privacy, detecting and preventing function creep (i.e., safeguarding data usage), and operationalizing the Findable, Accessible, Interoperable and Reusable (FAIR) principles for the data.

In this paper, we discuss the main data governance functionalities that are needed for data quality management and trust establishment within a smart city data ecosystem. Further, we elaborate upon a few emerging organizational aspects that are required for a good data governance regarding these functionalities. We consider these functionalities and organizational aspects because they are the main contributors of making tradeoffs among contending values in a smart city ecosystem. Realizing these functionalities indicates the need to exposing and ordering the issues that are involved in *optimizing data utility*, given that *privacy and ethical laws, regulations or guidelines are appropriately enforced, and that the data are efficiently and securely transferred, stored and organized*. To optimize data utility, important issues that should be considered are the quality of data and a proper implementation of the FAIR principles. The enforcement of ethic, privacy and security related laws, regulations or guidelines contributes to establishing and keeping trust in the data ecosystem. This requires establishing a sound mix of ethics, privacy and security measures. Note that in this contribution we explain the data governance challenges pertained to data quality as well as security and privacy. The issues that pertain to an efficient transfer, storage and organization of data are mostly technical in nature and are extensively discussed in computer science literature. Therefore, these issues are out of the scope of this paper.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss the concepts of smart cities and the role of data governance in smart cities. In Section 3, we present the key functionalities of data governance. Then, in Section 4, we discuss several emerging organizational aspects that should be considered for a proper implementation of the key data governance functionalities, i.e., the ingredients of an appropriate data governance. In Section 5 we elaborate on the first steps towards realizing the concept of smart cities. Finally, in Section 6 we draw some conclusions.

2. Background: Smart cities and data governance

2.1. Smart cities

Due to current rapid urbanization worldwide, cities have a huge impact on every aspect of our lives. In the last three decades, therefore, the concept of smart cities has emerged as a possible solution direction to address the problems stemming from this rapid urbanization [1,4,27]. The range of these problems spans from sustainability-oriented ones to efficiency ones [50]. Sustainability problems are related to social values (e.g., equity, community autonomy, citizen well-being, quality of life and gratification of fundamental human needs), economy vitality and diversity, and environment conservation (e.g., flora, fauna, and natural resources). Efficiency-oriented problems are related to efficient management of urban processes like transportation, education, and administration.

Consequently, the interest in smart cities has risen considerably among scientists, practitioners, and public policymakers in recent years. Despite this popularity, there is a lack of conceptual clarity around the term of smart cities due to the plethora of existing definitions [4,50]. According to [1], lack of a general agreement about the term *smart cities* may be attributed to the fact that the term has been applied to two different kinds of domains, namely hard and soft ones. In hard domains (such as, buildings, energy grids, natural resources, water management, waste management, mobility, and logistics), the Information and Communication Technology (ICT) plays a decisive role in the functions of the systems while in soft domains (such as education, culture, policy innovations, social inclusion, and government) the application of ICT is not usually decisive. The approach of hard domains aims at harnessing technology, particularly the ICT, to create a city that is, according to [29], instrumented (i.e., for capturing and integrating real-world data), interconnected (i.e., for communicating the resulting information among various city services) and intelligent (i.e., supporting operational decisions within city). The approach of soft domains sees smart city far from being limited to the application of technologies to cities as ICT eschews the actual knowledge about how cities function and disregards its complexity. This approach, instead, focuses on people and promotes nurturing citizens' creativity and knowledge (via education, culture and art) as well as their participation in order to achieve a sustainable growth.

Despite lack of a unified definition of smart cities, ICT plays a major role in shaping the concept of smart cities according to most definitions of smart cities, for an overview refer to [1,4,27,30]. According to the findings of [2, Section 4], the standardization bodies such as ISO, the ITU, the UK Standards BSI, and the US NIST have defined smart cities as "innovation – not necessarily but mainly based on the ICT –, which aims to enhance urban life in terms of people, economy, government, mobility, living and environment". For smart cities, ICT provides an infrastructure (and means) for collecting data from various data resources of a city, integrating the collected data into useful information, communicating the resulting information among city services, and supporting decision makers with collective intelligence that is based on the exchanged information [17].

As the collected data in a smart city impact its citizens on every scale, we believe that the ICT systems that collect and capitalize on them should not be perceived as pure technological systems. The way that data are collected, which are often blended with sensitive information about individuals, groups, and businesses, and the way that data driven ICT systems are devised, designed, implemented, deployed, and (mis)used, are going to impact us deeply at both individual and societal levels. As such, social transformations derived from data-driven systems must be attentive and respectful of the principles of social justice and ethics as postulated as the outer and inner boundaries of all human activity. Therefore, we are going to elaborate on our vision about how data in a smart city must be managed in a responsible way, as there are many data related blind spots like data quality issues and threats (e.g., "forms of dataveillance, social sorting and redlining, predictive profiling and anticipatory governance, nudge and [behavioral] change, control creep, and system security") in smart cities rhetoric [31, pp. 12].

2.2. A reference model for smart cities

Given that ICT is a main enabler of the concept of smart cities, it is useful to sketch a unifying model or architecture of smart cities to capture an abstract view of the complex context where ICT systems, in general, and data governance, in particular, operate. One interesting reference model for smart cities is presented in [46], which we

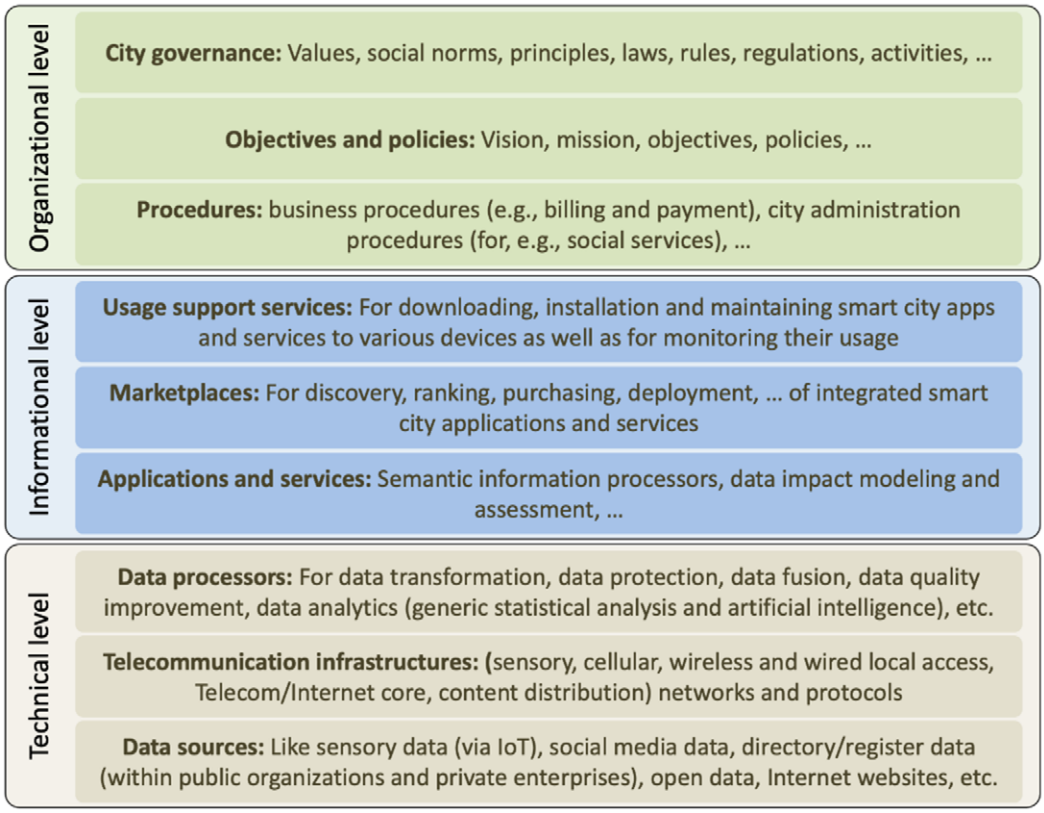

Fig. 1. A reference model for smart cities, adopted with adaption from [46].

have adopted with some adaption below for our purpose. The reference model, sketched in Fig. 1, comprises three levels, namely: organizational level, informational level, and technical level.

Within the technical level there are various data sources, telecommunication infrastructures, and data processors. Data sources, which provide raw data as well as (semi)processed data, include Internet of Things (IoT) sensors, social media, directories and registers within public and private organizations and enterprises, open data initiatives, and Internet websites. These data are offered commercially, openly (i.e., being public, free of charge and for any usage as in case of open data) or conditionally (e.g., for a given purpose by specific data consumer groups). Telecommunication infrastructures encompasses data communication networks of various types (like sensory networks, cellular networks, wireless and wired local access networks, and Telecom/Internet core networks) as well as telecommunication protocols to interconnect various devices in the smart city ecosystem. Data processors adapt and enrich heterogeneous data (by means of advanced data analytics, statistics and artificial intelligence) to produce information to the components of the informational level.

In the informational level there are various applications and services, marketplaces, and usage support services. Via applications and services, the processed data in the technical level are further enriched to support the semantic relations and understandings, needed for applying the resulting information in a smart city application setting. These applications range from those for critical systems (e.g., traffic light regulation and energy distribution systems) to those for infomercial services (e.g., about city traffic peak hours and citizen energy consumption). Marketplaces are online platforms enabling citizens and organizations to discover, rank, purchase, and deploy integrated smart city applications and services in different city domains. Usage support services help the end-users of smart city applications and services to download the apps to their devices (like smartphones, tablets, personal computers, and notebooks), to maintain these services, and/or to monitor the way that these services are used. Note that the end-users can be citizens, journalists, companies, or even other automated systems (like actuators to control physical systems in the smart city).

In the organizational level there are procedures, objectives and policies, and city governance. The procedures include the business procedures for commercial enterprises (like billing and payment) and administration procedures for (local) governments (like smart social services), which are partially enabled with (advanced) ICT and data-driven applications. The objectives and policies refer to the business and city governance objectives and policies that are based on business and city administration models, missions and visions and that are used to define, design and shape the above mentioned (business and city administration) procedures. The city governance refers to the values, social norms, principles, laws, rules, regulations and activities that guide the behaviors of the stakeholders (e.g., civil servants and citizens) and organizations towards defining and realizing the above mentioned procedures, objectives and policies. As such, city governance includes all organizational aspects of city governance.

In the next section we discuss the role of ICT resources in the reference model for smart cities.

2.3. Data governance

To achieve the sustainability and efficiency objectives of smart cities, we need to manage the ICT resources embedded in the informational and technical levels of the reference model (see Fig. 1) within and across the various organizations of a smart city ecosystem. For managing these ICT resources, as indicated in Fig. 2, we distinguish between governance and management, where the latter can be seen as the technical implementation of the former [23]. Two main categories of ICT resources are Information Technology (IT) resources and data resources. IT resources are the hardware and software components needed for collecting, storing and sharing smart city data. Management of these IT resources involves, among others, their installation, configuration, monitoring and maintenance and is mainly done within every organization in the ecosystem of smart cities with minor cross organizational coordination. Management of data resources, however, demands much more cross organizational collaboration and coordination relatively, due to the inherent property of data in traversing organizational boundaries. The difference in cross organizational natures of IT resource management and data management is indicated by the heights of the corresponding building blocks in the smart city reference model of Fig. 2.

In this contribution, we are going to describe our vision on data governance and, as such, our scope is beyond specific organization, excludes the IT resource management and IT resource governance aspects, and does not cover

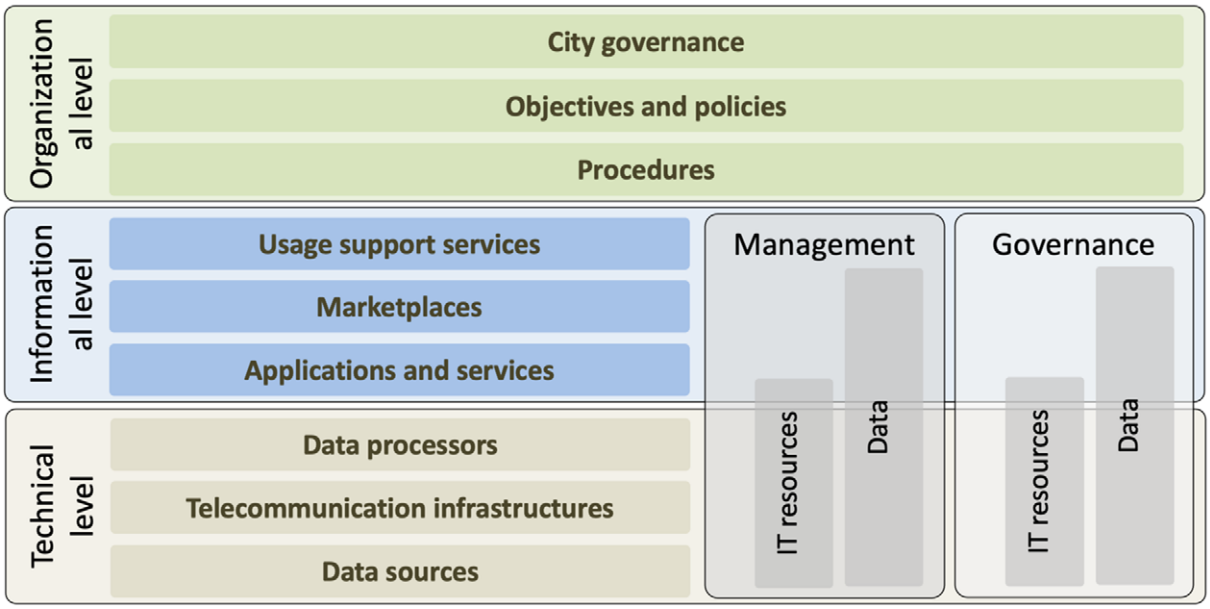

Fig. 2. The reference model of smart cities, extended with governance and management components.

the implementational aspects (i.e., it excludes the data management aspect).[1] In [12, pp. 432], *data governance* is defined as "the exercise of authority and control (planning, monitoring and enforcement) over the management of data assets". Data governance aims at defining the organizational structures, data owners, policies, rules, processes, business terms, and metrics for the whole lifecycle of data (i.e., collection, storage, use, protection, archiving, and deletion [23]. Cuno et al. [20] relate data governance in urban settings to identifying the entities who are involved in the management, provisioning and utilization of urban data (like data owners, data stewards, IT system administrators, data policy coordinators and data policy officers), and the required communication and interactions among these entities. These aspects (i.e., which entities and their interactions) are embedded in the decision-making process for data quality management, data access management, general data lifecycle management, and metadata management for the datasets in an urban environment.

In [19], the governance of smart city data is viewed more broadly than just focusing on the substance of rules around the collection, use, sharing, retention, and disposal of data. Their suggestion is to focus on how these rules are made, disputed, and changed (i.e., on who is making the rules and what processes are used in rule making). To this end, the authors adapt four principles of Ostrom's 8 principles that are applicable to the context of smart city data. These procedural principles [19, pp. 7] are:

1. "Promote responsibility for data governance among multiple layers of nested enterprises" to collaborate or make collective decisions,
2. "Create processes for the members of the affected community" (e.g., citizens and civil servants) "to participate in making and modifying the rules around data", aligned with the community's self-identified needs,
3. "Develop an effective monitoring system to be carried out by the community" members, non-governmental organizations, watchdog groups, media outlets, independent oversight bodies, or industry standard working groups, and
4. "Provide accessible means for dispute resolution, use graduated sanctions against rule breakers, and make enforcement measures clear".

Eke and Ebohon [22, Section 3.2] consider a data governance that "incorporates data from all relevant sources, focuses on diverse values for all stakeholders and aligns with strategic objectives for and by residents". As such, this view pays special attention to the impacts of the inferential decisions made with the data (due to, for example, exclusion of some groups in the population). This view extends the scope of data governance from "ensuring maximum value creation from the data while adhering to ethical and legal requirements" to "the establishment of processes, policies, roles and responsibilities that foster the effective management of data for the benefit of relevant stakeholders that will be affected by the decisions derived from the data" [22, Section 3.1]. To this end, smart cities should set up a data governance framework with the right data (i.e., data without representation bias, measurement bias, evaluation bias, aggregation bias, population bias, sampling bias or data linking bias), the right algorithms (i.e., in being accountable, explainable and fair), the right people (i.e., the right mix of people and expertise to adequately address inclusion issues), and the right policies/standards to build trust, transparency, accountability and responsibility) [22]. This is an interesting viewpoint as it asks for monitoring the impacts of data when the results of data-driven applications are applied into practice. In other words, managing the use of smart city data belongs to the domain of data governance in addition to managing the quality and diverse nature of the data that inform data driven applications in a smart city eco-system.

In Section 3, we discuss several functionalities for managing the quality and diverse nature of data. Section 4 is devoted to a set of building blocks for managing the use of smart city data.

3. Data governance functionality

In smart cities various types and enormous amount of data are collected and fed to data-driven smart city applications. These data originate from various sources (e.g., IoT devices, wearables, personal mobile devices, drones

[1] Note that managing the IT resources and managing the data resources are not entirely independent. Further note that "the distinction between IT governance and data governance is currently not clearly established in smart city initiatives" [34, pp. 11].

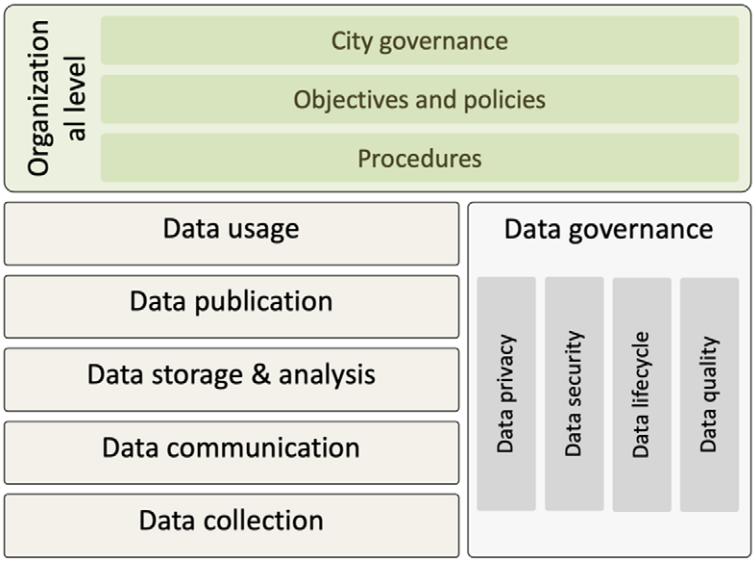

Fig. 3. An illustration of data governance functionality relative to the smart city reference model shown in Fig. 2.

and robots, and the data registers/databases within public organizations and commercial enterprises) with varying data quality levels. Often these data are (or can be) related to individuals and, as such, they might be personal data or even sensitive personal data (like someone's gender and health status). Non-personal data may still be sensitive due to conveying, for example, sensitive information about businesses or public safety affairs. Moreover, smart city applications (like smart social services, smart transportation services, e-government services, public health services, predictive policing, public safety services, and emergency services) deliver added value services for citizens, governmental organizations, and commercial enterprises with huge impacts on individuals, economy and society.

Data governance in smart cities should aim at improving the quality of data to foster the desired impacts of smart city applications. Concurrently, data governance in smart cities should also aim at containing the undesired impacts of these applications and/or of the collected sensitive (personal or business) data. Containing these undesired impacts enhances the trust of stakeholders in the approach of smart cities and consequently fosters the growth of smart cities. In this section we elaborate on improving the data quality of smart city (see Section 3.1), while enhancing the trust of stakeholders (see Section 3.2), and finally recapture the concept of smart city governance (see Section 3.3). To illustrate data governance in smart cities, Fig. 3 shows a modified version of the reference model in Fig. 2 with its data quality and trust aspects as well as with the data analytic stages (i.e., data collection, data communication, data storage & analysis, data publication and data usage) that are relevant for data governance.

3.1. Addressing data quality issues

An important factor that should be considered is to establish an acceptable level of data quality. A bad quality of data often results in a low utility of the data since the data will be considered as unsuitable for use. Therefore, it is necessary that the data meet a standard quality level. However, to predefine an acceptable data quality level is a tough task. As will be argued, this is because the notion of data quality is hard to capture and an acceptable data quality level depends on the type of the application and its deployment context.

The broadness of the notion of data quality makes it hard to capture in a single definition. Therefore, there is no widely accepted definition for data quality. Various definitions of data quality can be found in literature, ranging from defining some data quality dimensions (such as accuracy, completeness, timeliness, usability, relevancy and reliability) to more comprehensive (i.e., generic) definitions [26,47,51,52]. Some of these dimensions, e.g., accuracy, may be classified as objective while the other dimensions can be seen as subjective, e.g., relevance. This holds also for the comprehensive definitions of data quality. For example, an objective and comprehensive definition is that

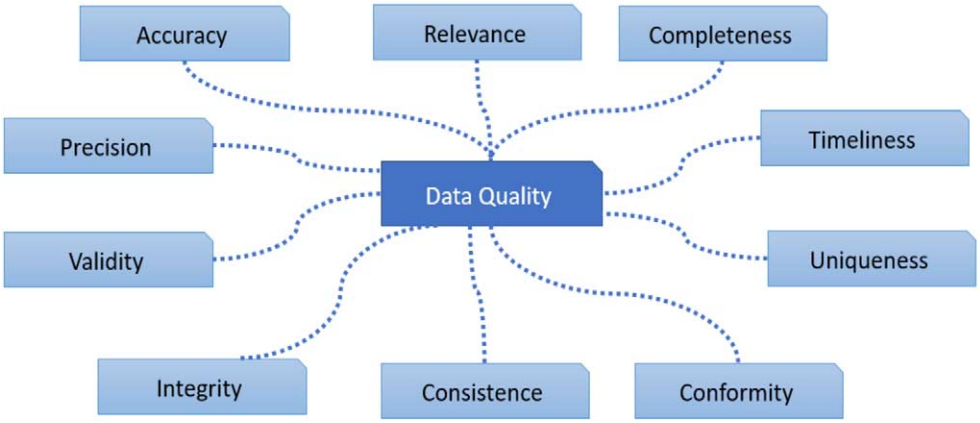

Fig. 4. Some data quality dimensions.

data should adequately represent (parts of) the real world, while a subjective definition is that data should fit the usage purpose of an application. Although the definitions of data quality may look different, the comprehensive definitions subsume the core elements of the more specific definitions.

Assessing the degree to which a data set represents the real-world is a predominantly objective definition of data quality. As elaborated in [51], this implies that there should be a valid and unambiguous mapping from the data set to a complete description of a real-world phenomenon. Although this definition raises a number of questions, such as how to determine the completeness and which real-world phenomenon should be represented (e.g., those of today or the past), it provides sufficient guidelines for operationalization. Based on this definition, scholars have specified several data quality dimensions such as data completeness and accuracy. Assessing the degree to which a data set fits its usage is a predominantly subjective definition of data quality. Also this definition raises several questions such as to which usage the data should fit and how this fitness should be defined. Elaborating upon this definition, scholars have specified several data quality dimensions such as data relevance, usability and understandability. These dimensions, which depend on the type of application at hand, provide more criteria for operationalization of the data quality concept.

Thus, from data quality definitions a wide range of subjective and objective dimensions can be derived. In Fig. 4, we depict some of these dimensions, see [11] for a detailed overview.

Alternatively, in [42] a semiotic view is taken on data quality, which results in an analytic framework for data quality. Three analysis levels have been distinguished in this framework, namely: the syntactical level, the semantical level and the pragmatic level. The syntactical level focuses on the degree of adherence of the data to predefined meta data characteristics. As such, it measures the consistency of the data within a data set. The semantic level focuses on the meaning of data, for example, whether it is complete, and up to date. The pragmatic level is concerned with the utility of data in terms of, for example, usefulness and usability. On this level the assessment of data quality depends on the involved professionals, the application(s) at hand, and the organizational context of the use.

In order to make the concept of data quality tangible, we combine the semiotic view with the data quality dimensions, resulting in a data quality framework. This framework can be represented by a matrix as depicted in Table 1. Each dimension, which constitutes a column of the matrix, may be decomposed into a set of indicators. An indicator pertains to one of the levels, which constitute the rows of the matrix.

The following example illustrates how each dimension is decomposed into a set of measurable indicators and how the appropriate level is determined for each indicator.

Example. Consider the data quality dimension of completeness. Completeness may be decomposed into two meaningful indicators, namely: the number of NULL values and the existence of missing relationships. The former typically pertains to the syntactical level, while the latter pertains to the semantical level. As explained in the following, NULL values in a database complicate the interpretation of database results. Therefore, strategies to reduce these values increase the data quality.

Table 1
Data quality matrix

		DQ dimensions			
		Completeness	Accuracy	...	Relevance
Levels	Syntactical	The # of NULL values	-		-
	Semantical	Missing relationships	-		-
	Pragmatic	-	-		-

Table 2
An example of a database

Tid	Age	Gender	Category	Price	Damage
100	20	Male	Leased	70K	Yes
200	35	Null	Not leased	80K	Yes
300	24	Female	Leased	75K	Yes
400	28	Male	Not leased	40K	Yes
555	28	Male	Leased	50K	No

To illustrate the impact of NULL values, consider the database in Table 2, referred to as *damage database*, which consists of 6 attributes and 5 tuples. Note that the attribute tid stands for tuple identifier and uniquely defines a tuple. Further, suppose that the following constraint is defined on the database: the number of persons = the number of females + the number of males. Now we pose the following question: How many persons are in the databases? An optimizer of a database management system chooses its own strategy to execute a question according to a query execution plan. In this case, we consider two options to answer the question, each leading to different answers. The first option is to count the number of tids, leading to the answer 5. The second option is to apply the constraint that the number of persons = the number of females + the number of males, leading to the answer 4. At a first glimpse, one may choose for the first option since there are indeed 5 persons recorded in the relation. However, this answer may cause confusion if a subsequent question is to split the persons along gender. The answer will be then 3 males and 1 female, which does not sum up to 5.

To illustrate a missing relationship as a semantical incompleteness, consider the following example where some queries cannot be processed. Suppose that we pose the following query to our damage database: What type of cars, which are leased by males between 20 and 30 years old, are involved in damages? Such a query cannot be answered (even if we have a separate database with types of cars), since the relationship between the damage database and types of cars has not been established. Therefore, at a semantical level the database can be marked as incomplete to a certain degree.

As stated before and in line with [32,35], we believe that the assessment of data quality is eventually dependent on the applications at hand. Consequently, knowledge about the applications and their situational context, referred to as domain knowledge, is of vital importance to assess and improve the quality of data. In turn, the domain knowledge determines the data quality dimensions that should be selected and the proper levels on which the data quality can be assessed. From a practical point of view, we foresee that a crucial task of a data ecosystem is to come up with typical sets of data intensive applications for an application domain. For each set of applications, the proper data quality dimensions, the corresponding data quality indicators and the proper level(s) for each data quality indicator should be determined. This will result into a filled data quality matrix as depicted in Table 1, and for each element in the matrix an (acceptable) value should be specified. We refer to such a filled matrix as an instantiation of the data quality matrix. The values assigned to the elements in the data quality matrix may represent the minimum or desired values for an indicator at a certain level that should be met for a desired quality of the data. Actually, an instantiation represents a typical scenario within a domain, i.e., represents a marketing scenario. In a marketing scenario, city marketers need data to promote the city. Therefore, the granularity/accuracy of data (i.e., an exact spelling of the hotspots in a city) will be less important for the quality of the data.

The challenge for every application domain is to come up with a number of useful data quality instantiations.

3.2. Addressing trust issues

In addition to addressing data quality issues, the success of smart cities depends on addressing the trust issues that may arise due to security, privacy and ethics related threats. Collecting, communicating, storing, processing and using (privacy) sensitive data should respect the required and expected level of security, privacy and ethics in a (loosely coupled) data ecosystem. Otherwise, the threats and risks introduced to privacy, human rights, and ethical or democratic values may not only hamper the success of smart cities but also inflict harmful and irreparable impacts on citizens and society. In this section, we elaborate on privacy governance (Section 3.2.1) and security governance (Section 3.2.2), where the latter is done in a less elaborative way than the former. Due to space limitations, we leave out the governance of ethical aspects here, noting that some of its characteristics are common with those of privacy and security governance.

3.2.1. Privacy governance

Privacy is a normative concept that is deeply rooted in various disciplines such as philosophy, law, ethics, politics and sociology [40]. Many efforts have been put to conceptualize privacy (i.e., to define what privacy is, what makes it unique and distinct) by searching for a common set of necessary and sufficient elements that single out privacy as unique from other conceptions [48]. Solove [48] analyzes a number of these definitions and argues that these privacy concepts are either over inclusive (i.e., being too vague) or too restrictive (i.e., being too specific). Subsequently, Solove concludes that privacy cannot be conceptualized in a definition with some necessary and sufficient conditions (i.e., based on inclusion and/or exclusion rules). Instead, Solove proposes a framework for personal data protection, where one should identify and deal with possible privacy risks in various stages of the data analytics process. As indicated in Fig. 3, these stages can be data collection, data communication, data processing and storage, data publishing, and data usage. Similarly to [25], we define *privacy governance* as a subset of corporate governance, which primarily focuses on establishing and maintaining the control environment to manage the risks relating to the possible privacy breaches of personal information assets during various stages of the data analytics process. Privacy governance is the combination of technical, organizational and regulatory approaches for the governance of privacy [28].

Current privacy governance instruments (like personal data protection regulations, directives, guidelines and standards) mostly predate the current smart city developments where digital technology has become pervasive and a new generation of users has grown [49]. More specifically, nowadays we witness that the amounts and types of available data are enormously increased (consider the rise of the big data paradigm), the number of data-driven applications has increased (consider the emergence of many things of smart things), the individuals who use data-centric technologies have become active information consumers (consider the daily use of search engines by common citizens) and/or active information producers (consider the explosive growth of user generated content and social media content). Tene [49] mentions several consequences of the recent developments (i.e., the pervasiveness of new digital technologies and the emerging behavior of its new generation users) as follows:

- The distinction between personal and non-personal data becoming fuzzy and muddled,
- The identifiability aspect of the definition of personal data becoming obsolete due to the ability of behavioral targeting without deciphering specific identities (i.e., attribution, see [8,9]),
- The distinction between what is public and private is eroding,
- The notice and choice paradigm (in the United States) and the transparency principle (in the EU) has become meaningless because the current lengthy and complex privacy policies have little or no use to users,
- The allocation of obligations (and accountabilities) among data controllers and data processors becoming distressed due to the increasingly collaborative manner in which (businesses) partners operate nowadays,
- The assignment of the data controller role in social media becoming difficult and unclear (i.e., it is the user who posts information voluntarily and makes privacy choices or the social network service which stores the data, processes and secures it, and sets privacy defaults),
- The inability to forget past transgressions from the personal data collected and retained by various parties, and
- The existing requirements and restrictions on global data transfers being unrealistic, bureaucratic, costly and ineffective.

Consequently, for example, there is currently no global actor to actually enforce data protection in global cyberspace [21] or the current (self) regulatory mechanism of notice and consent could not be enforced effectively and informedly (i.e., in way that data subjects understand the risks of processing their personal data) in the era of big data because the inductive power of data analytics implies looking for and finding new purposes if we want to realize the promise of the technology [49], where the new purpose (i.e., the secondary uses) was unimaginable when the data were collected. This (generation of) secondary data usage may lead to function creep to be discussed in Section 4.3. As such, the current (self) regulatory mechanism of notice and consent should be reconsidered (or maybe even discarded) in the era of big data analytics.

Nowadays, personal data protection "isn't merely about excluding or banning some person or information from all public contact or disclosure, but rather the management and fine-tuning of an individual's exposition to the world" [21, pp. 52]. Therefore, there is a need for a shift in the theoretical and conceptual underpinning of privacy governance, especially in the context of smart cities. To this end, Raab [43] suggests and foresees four interrelated developments in privacy governance, namely: the rise of accountability as a regulatory and self-regulatory philosophy and technique, the (re-)discovery of ethics and its potential shaping of principles, rules and behavior in the processing of personal data, the increased prevalence of risk- and harm-based understandings of privacy regulation, and the conceptualization of privacy as having a social value as well as being an individual right.

All this puts several challenges in front of us in realizing a sound privacy governance for smart cities. A fundamental reshuffle of the current privacy governance asks for a privacy governance framework that (from [49]):

- is attuned to identifying and mitigating privacy risks, rather than making a dichotomy between personal and non-personal data, or between private and public spheres,
- provides a new approach to notice and choice, with an emphasis on enhancing user awareness and understanding rather than presenting corporate disclaimers of liability,
- allocates responsibility according to data usage and the risks inflicted to data subjects, rather than making a formal dichotomy between data controllers and data processors,
- makes a sensible balance between data retention needs and individuals' right to be forgotten; and
- governs cross border data transfers based on accountability and ongoing responsibility, rather than creating arbitrary barriers and requiring bureaucratic form fillings.

Privacy protection has been often framed in an individualistic perspective (e.g., in being a private good that affects one's own self as captured by most privacy laws and regulations). However, privacy protection has increasingly become enacted in complex socio-technical systems. For example, the contextual integrity theory of [39] aims at capturing the social nature of privacy for information flows in their social context (i.e., requiring compliance with contextual informational norms for these flows). However, the contextual integrity theory does not elaborate on how the contextual informational norms can be established in a social context. To this end, the Governing Knowledge *Commons* framework can provide a way to derive these contextual informational norms [36].

The knowledge commons framework envisions actors beyond individuals being in a broadly defined social context (such as education and healthcare) and regards them as members of a community involved in producing or managing a set of resources. To this end, the actors (co)produce the applicable rules-in-use for managing these resources (i.e., being involved in the governance). Rules-in-use are the actual rules and different from nominal rules (i.e., those in the books). Rules-in-use emerges (perhaps unanticipated) in practice from "interactions within often complex structures of formal and informal institutional arrangements" [45, pp. 118]. Rules-in-use include various sorts of constraints such as nominal rules, social norms, strategies and tactics of compliance and avoidance, power dynamics, and enforcement mechanisms. It is interesting to investigate the relationships between contextual informational norms and individuals' privacy expectations/preferences, and how/whether participation and community engagement can shape individuals' privacy expectations/preferences.

In summary, there is a need for alternative types of agreement such as the formation of multistakeholder bodies and mechanisms to help privacy governance in cyberspace [21]. We envision that design thinking is one of the promising methods for harvesting and exploring the multistakeholder approach and involving different actors in the process of data governance (see Section 5 as well as [5,7,10]). Furthermore, design thinking can be used to derive and assess the appropriateness of contextual information (i.e., rules-in-use) norms a well.

3.2.2. Security governance

Information security was evolved from the early field of computer security. Information security, according to the US Committee on National Security Systems (CNSS), is concerned with "protection of information and its critical elements, including the ICT systems (software and hardware) that use, store, and transmit that information" [53, pp. 8]. Information security aims at protecting several so-called *critical characteristics of information assets*, whether in storage, processing, or transmission. The traditional critical characteristics of information assets that should be protected are *confidentiality* (to protect information from disclosure or exposure to unauthorized individuals or systems), *integrity* (to protect information so that it is whole, complete, and uncorrupted), and *availability* (to enable authorized entities to access information without interference or obstruction, and with the required data quality). Information security is achieved via the application of policy, education, training and awareness, and technology. These traditional information characteristics to be protected via cybersecurity are referred to as the *CIA triangle*, where the abbreviation refers to their initial letters. According to [25], information security governance is a subset of corporate governance, with its primary focus on establishing and maintaining the control environment to manage the risks relating to the CIA characteristics of information assets.

The concept of smart cities encompasses an Information System (IS) that is ubiquitous and distributed with loosely coupled subsystems, administrated by various organizations (i.e., administrative domains). Security governance in smart cities should aim at creating a trusted computing environment that supports safe and secure collaboration within and among the participating entities. This is a challenging task, particularly in the loosely coupled setting of smart cities (i.e., in open or partially open smart environments, across multiple organizations), where one should provision secure interactions between parties who do not necessarily have a common security infrastructure for sharing information assets reliably. Organizations in smart cities administrate the security of their ICT infrastructures differently, have various devices and computing machines with varying processing and memory capabilities, and above all, pursue different missions and objectives in a data ecosystem.

Due to space limitation, we do not plunge into particularities of security governance here; also, because similar privacy governance challenges (as mentioned in Section 3.2.1) apply to security governance in smart cities. It is, however, worthwhile to highlight the dependencies between privacy protection and cybersecurity. All experts and people know that privacy protection requires establishing cybersecurity. Nevertheless, the other way around is not common knowledge. Information sharing is one of the pillars of cybersecurity, especially in distributed settings such as the Internet itself, IoT systems, and the distributed Intrusion Detection Systems (IDSs) used for detecting cyberattacks. The data collected by local entities (e.g., organizations) are often privacy (and business) sensitive. Therefore, sharing such information across organizations is not self-evident as it may compromise the privacy of individuals, the competitive advantages of businesses and the national sovereignty of countries. Therefore, it is a challenge to determine what (aggregated) information is relevant to share among organizations in smart cities. To this end, an appropriate security governance should consider two types of privacy:

- The privacy of the individuals within organizations, who are victims of cyber attacks (i.e., the privacy of victims). Organizations may not be willing to participate and share information if this information is privacy sensitive.
- The privacy of the individuals being suspicious as cyber attacker (i.e., the privacy of suspects). For sharing information about the possible cyberattacks it is necessary to share some personal data (like IP-addresses) of potential attackers in order to locate them effectively. As such data sharing, if done inappropriately, may lead to imposing sanctions against alleged, but not proven, cyber attackers.

Often the operational environment of smart city applications/systems is dynamic (i.e., participating individuals, devices and resources may join and leave at any time), distributed (i.e., the physical boundaries of the environment may span over many locations), and loosely coupled (i.e., the participating bodies do not belong to a single organization's administrative domain). These characteristics inflict heavy burden on securing smart city systems/applications. The traditional security paradigms that are based on dividing a system domain into trusted and untrusted sub-domains (where trusted insiders and untrusted outsiders, respectively, reside) do not work well for such smart city applications/systems as they often have no clear network boundary to separate insiders and outsiders. Recently a new cybersecurity paradigm called zero-trust has emerged as promising to protect such ISs. The Dutch National Cyber Security Centre in a recent report has advised organizations to deploy the zero-trust model in their future

investments [38]. For IoT based systems there are also a rising number of publications advocating the zero-trust model to protect these systems [14,41].

According to the zero-trust model, one should never trust anybody, either being inside or outside a network, until the person is verified [3]. This model, which assumes that both internal and external networks cannot be trusted, asks for a more granular ruleset to protect ISs. It presents more an attitude towards system protection than a system architecture [3]. For example, in some situations the zero-trust model may boil down to ensuring that data are securely accessed based on the identity and contextual situations of users, with strong access control mechanisms in place, and with registering data usage logs. It may also require a segmentation of networking and data processing resources so that users' access can be limited to those segments that are strictly necessary for provisioning the service in mind, and nothing more. Note that zero-trust is more than just network segmentation, "zero trust focuses on protecting resources (assets, services, workflows, network accounts, etc.), not network segments, as the network location is no longer seen as the prime component to the security posture of the resource" [44, pp. ii].

The basic tenets of a zero trust architecture, according to [44], are: (1) All data sources and computing services are considered resources, (2) all communication is secured regardless of network location, (3) access to individual enterprise resources is granted on a per-session basis, (4) access to resources is determined by dynamic policy, including various behavioral and environmental attributes, (5) the enterprise monitors and measures the integrity and security posture of all its assets, (6) all resource authentication and authorization are dynamic and strictly enforced before access is allowed, (7) the enterprise collects as much information as possible about the current state of assets, network infrastructure and communications and uses it to improve its security posture.

Although the zero-trust model presents a promising model for dealing with security issues of new ISs such as smart city applications, it may inflict several challenges on organizations adopting the model. During transition to the zero-trust model, organizations may face new threats due to the security gaps created if the transition is realized in several (small) steps [44]. Further, lack of commitment to its realization well may become an obstacle to its success and it may result in lack of productivity if users are wrongfully locked out of the resources necessary for their well-functioning [44].

3.3. Recapturing the concept of data governance

Like [12], we consider data governance as the exercise of authority and control (planning, monitoring and enforcement) over the management of data assets. As such, data governance aims at defining the organizational structures, identifying the actors (like data owners, data stewards, IT system administrators, data policy coordinators and data policy officers), defining the policies and rules, designing the processes and business terms, and specifying the metrics that are relevant for the whole lifecycle of data (i.e., collection, communication, storage and processing, dissemination and use). Note that

- We distinguish between data governance and data management, where the latter is the technical implementation of the former.
- Like [19], we perceive the scope of smart city data governance more than just the substance of rules around the data lifecycle and include also *how* these rules are made, disputed, and changed (i.e., on who is making the rules and what processes are used in rule making).
- Like [22, Section 3.1], we consider the scope of data governance being extended from "ensuring maximum value creation from available data while adhering to ethical and legal requirements" to "the establishment of processes, policies, roles and responsibilities that foster the effective management of data for the benefit of relevant stakeholders that will be affected by the decisions derived from the data". This extension asks for proactively monitoring the impacts of data when the results of data-driven applications are applied into practice.

We elaborated on two main aspects of data governance: Improving the data quality (Section 3.1) and enhancing the trust of stakeholders (Section 3.2). This elaboration and the data governance model shown in Fig. 3 are inspired by the study of [12].

4. Organizational aspects

Once the functionalities of data governance have been determined, a proper organizational structure of the smart city data ecosystem needs to be set up to execute these functionalities. In this section, we discuss several issues that should be considered in determining an organizational structure. First, in Section 4.1, we discuss the entities that should be involved in a smart city data ecosystem. As discussed in the previous section, data governance encompasses a broad range of complex and advanced functionalities. Therefore, a successful execution of these functionalities asks for having data governance knowledge from different areas of expertise that can hardly be found at a single entity in the data ecosystem. In Section 4.2, we focus on the dissemination of data governance knowledge across the entities in a data ecosystem. Finally, Section 4.3 is devoted to the FAIR principle, where we especially stress on the issues involved in the re-usability of data in a responsible way (via, for example, preventing function creep, having usage control, establishing interoperability, and creating data map and metadata).

4.1. Data domain

A data domain refers to the *set of organizations (including their human capitals and technical infrastructures)* that serve a specific (set of) data-driven application(s). Every data domain may have a data ecosystem among the participating organizations for managing its data. To determine which organizations should be part of a data domain (or its data ecosystem), it is worthwhile to identify the organizations that perform the tasks related to the data of the data domain. These data related tasks can be broad but roughly we distinguish the following main data related tasks (see also Fig. 3):

(1) Data collection: To collect data from, for example, embedded systems and individuals,
(2) Data transfer: To exchange data via communication networks and systems such as WiFi networks and Internet gateways,
(3) Data storage and analysis: To process data by using, for example, big data technologies,
(4) Data publication: To share and disseminate the processed data together with the corresponding metadata,
(5) Data usage: To use the data by a single unified (automated) system,
(6) Data governance: To efficiently and effectively govern the way that the data of the data ecosystem are managed.

The organizations that are involved in performing at least one of these tasks are eligible for participation in the data ecosystem and are members of the corresponding data domain.

The organizational structuring of a data domain (i.e., the structure of the relationships of the organizations in a data ecosystem) depends on the stakeholders involved and the generic domain in which the data ecosystem is embedded. At a high abstraction level, a generic *domain* may be considered as a set of organizations that serve a specific aspect of our society such as healthcare, science, justice, public safety and education. The organizations in a generic domain are coupled strongly or loosely. In other words, these organizations have already an existing formal or informal structure for collaboration, like a chain structure, peer to peer structure or a hierarchical structure. The hierarchical structure is mainly found within an organization (i.e., among its various departments), while the other structures are often found across organizations. For example, the supply chain within the logistics domain has a chain structure typically, while collaboration in the science domain has a peer-to-peer structure generally. In a local government (like that of a city) or a central government several variants of the chain structure can be recognized. The organizational structuring of a data domain, generally, inherits that of the corresponding generic domain.

To realize data governance within a data domain, several measures should be taken. Data governance realization may call for introducing some (new) roles within a data ecosystem such as data steward and (chief) data officer roles. Although their importance is being recognized and they have been introduced in practice slowly, these roles and their corresponding tasks are not well defined and stablished generally. An elaboration of these roles and their corresponding tasks should be on the future research agenda of smart cities. Moreover, for realization of data governance, we need, among others, to educate the participating organizations about the data governance and implement the key functional features of data governance. In Sections 4.2 and 4.3 we are going to elaborate on these data governance knowledge distribution and the key data governance functional features, respectively.

4.2. Knowledge distribution

Realization of data governance (and the data management) in practice asks for deploying data quality management mechanisms and data protection mechanisms (like data security and personal data anonymization mechanisms) within and across smart cities data domains. Often, these mechanisms are complex and therefore a successful deployment of smart cities asks for disseminating data governance knowledge among the involved parties so that they (i.e., citizens and organizations) who reside closely to their data can play their roles in maintaining the quality of their data as well as protecting their data.

The knowledge and expertise about (new) data governance (and data management) can be deployed within the data domains of a smart city in various ways. To start with, we investigate two structures that specify the two ends of the deployment spectrum, namely: fully-distributed deployment and centralized deployment. In the fully-distributed deployment, every organization in the data domain hosts the knowledge about data governance/management. This is the most desired format because the domain knowledge (i.e., the knowledge of the organization) can fully be exploited to adapt data quality to a desired level while protecting the raw data at the source (e.g., without letting the data leave the premises and boundary of the organization in charge of personal data collection). However, it is a serious challenge for every organization to master data governance (and data management) knowledge and carry out it independently and uniformly.

The second option centralizes the data governance (and data management) knowledge at one organization (i.e., at the central party). However, on the downside, the central party in this model must be trusted to receive the raw data of all local parties (i.e., it must be a Trusted Third Party, TTP). This option is challenging as, for example, the collection of privacy-sensitive data at a central party is a classical threat for privacy protection. Furthermore, the scalability of the data governance (and data management) functionality becomes an issue as the central party becomes a bottleneck with the current fast growth of data and data sharing. Table 3 summarizes the pros and cons of the fully distributed and fully centralized options.

Considering the pros and cons of the first two options mentioned above, our envisioned framework for disseminating data governance (and data management) knowledge in smart city settings is based on a third option that offers a partially distributed data governance (and management) to benefit from both distribution and centralization while avoiding their disadvantages as much as possible. In this way, member organizations do not become overwhelmed with demanding data governance (and management) tasks. For distribution of data governance (and management) knowledge between the central organization and the member organizations, we envision gradually transferring the data governance (and management) knowledge to the member organizations as much as possible. In this way, creating data governance (and management) expertise at member organizations occurs gradually, without imposing immediate burden on them to learn and apply complex and demanding data governance (and management) tasks. Initially, we foresee that the member organizations learn the fundamentals of data governance (and management) and learn an initial set of guidelines and apply them into practice.

There are several benefits of learning about data governance (and management) as much as possible. Firstly, data stewards at local parties can better understand the data quality issues and the (privacy, security and ethics related) risks associated with the data they collect, process and share. This understanding can be helpful to project own domain knowledge into data quality management and data protection. Secondly, the data stewards are often processors/consumers of external data sets, which are adapted (e.g., anonymized) by other parties. Knowing about the

Table 3

Table a summary of the pros and cons of the fully centralized and distributed configurations of data governance deployment (i.e., of data management)

Data management functionality	Pros	Cons
Fully distributed	+ Sensitive raw data remaining in their domains + Workload distribution	– Lack of enough data management resources at local parties – Lack of coordination for applying data management uniformly
Fully centralized	+ Establishing stronger data management at the central party + Fully coordinated data management functionality	– Sensitive raw data crossing their domain boundaries – Overload of the central unit – The central unit being the single point of failure

technical details of data governance (and data management), they can avoid misunderstandings when processing such data. Learning the initial set of guidelines can enable member organizations to carry out routine data governance (and management) tasks as much as possible. The central party, which comprises several data governance (and management) experts, can gradually educate member organizations and offer consultancy to them.

To overcome the barriers of adopting complex data governance (and data management) in organizations, we envision an evolutionary process to expand the initial data governance (and management) knowledge, conveyed from the central party to member organizations. The distributed knowledge can be of awareness, principle and how-to types as defined in [33]. These are the different types of innovation knowledge used during the technology adoption process for mobilizing potential adopters. The a*wareness knowledge* is about the existence and key properties of an innovation, the *how-to knowledge* is about how to use an innovation properly at individual and organizational levels, and the *principle knowledge* is about the functioning principles underlying the innovation.

4.3. FAIR principle

To promote re-usability aspect of the FAIR principle, it is necessary that data are findable and accessible. Furthermore, many applications demand that various data sets are integrated. Therefore, these data sets should be interoperable. Thus, re-usability requires the compliance of the findable, accessible and interoperable aspects of the FAIR principle. Good data governance should take care of the embedding of the FAIR principle among the organizations within a data domain (or data ecosystem).

There are many efforts and best practices reported in the literature to make data findable and accessible. Choices should be made within a data ecosystem about the best ways to make data findable and accessible within a data domain. Furthermore, it should be made clear which organizations in the data ecosystem are responsible for these aspects.

Interoperability is a more challenging topic and requires more coordination to address it properly. In general, the organizations involved in a data domain collect and store data sets to optimize their core business processes. Whether these data sets are interoperable with the data sets stored at other organizations is not their primary concern. Since interoperability is a necessary condition for data integration, it should be on the agenda of data governance. Standards to exchange data should be defined. Attributes that every data set should include for interoperability purposes, e.g., primary/foreign keys, should be defined as well. Finally, the quality standard that a data set should meet should be defined. As discussed in Section 3.1, data quality may be decomposed in different dimensions and on several levels.

For re-usability of data sets, the findability, accessibility and interoperability of data sets are necessary conditions, but they are not sufficient. To promote re-usability for every data set, it is necessary to clarify the semantics of the data, the purposes for which the data may be used, and the conditions of the data usage. There are several ways to realize this, for example, by specifying the metadata that provide insights in the data quality, and providing an unambiguous description of every attribute in the data set, giving insights about the type of analysis that the data is suitable for. All these mechanisms may be tailored for different types of users. Good data governance includes well-considered choices about these mechanisms and defines the responsibilities and tasks among the entities in the data ecosystem.

An important emerging side effect of re-usability is function creep. Function creep refers to the situation where data to the corresponding applications are gradually used differently than they originally were intended to be used. Facebook was once intended to connect people, but it is now also a tool for organizations to collect as much data as possible. It often starts with collecting microdata to provide services to users. Microdata concern the data of individuals, which consist of some attributes about them. However, these individuals and/or data providers have not full control on the operations that are performed on the data over time [6]. For example, the data can be combined with other data, then aggregated and used for (direct) marketing purposes. Aggregated data can also be used to formulate policies. The microdata can also be used for surveillance purposes in combination with other data. Function creep often takes place gradually and out of the sight of stakeholders [15]. As a result, it takes a while before stakeholders become aware that the data are used differently than that was initially agreed upon. If function creep becomes a frequent occurrence, especially when the stakeholders are harmfully affected, then the support for maintaining a data ecosystem will decrease. The impression may arise that function creep is a strategy to present stakeholders with a

fait accompli. Therefore, it is necessary to monitor function creep closely to prevent its occurrence. In particular, one should prevent using data for tracking and tracing (of any group) of individuals as a result of function creep. In the Netherlands, we are already observing an increasing resistance of citizens in using an app that is intended to control the spread of COVID-19 due to the fear of function creep. Finally, we note that function creep may lead to a violation of the privacy of citizens and to stigmatization of groups of individuals. We believe that an appropriate data governance should take care of function creep.

As illustrated in this section, the embedding and elaboration of the FAIR principle come with a number of challenges. Since the FAIR principle is considered as a cornerstone for data ecosystems, these challenges need to be systematically addressed in the future research.

5. Discussion: Towards realization of smart cities

In this section, we elaborate on two aspects of how data governance can be made, disputed, and changed. In Section 5.2 we explain the need for making trade-offs among contending aspects of data governance (i.e., between data quality versus trust establishment aspects) and in Section 5.3 we discuss the need for collaborations to make these trade-offs. We start the section with providing a brief overview of a methodology that can be instrumental for making the mentioned trade-offs and collaborations (see Section 5.1).

5.1. Design thinking

Design thinking is a methodology to express concerns, needs and wishes of stakeholders in an explicit manner and to create support and consensus among them. Design-thinking has shown to be useful in settings where user needs and concerns are insufficiently formulated and are hidden in tacit knowledge. In such settings, there are often different and poorly communicating stakeholders [5]. A typical design-thinking process comprises the following stages: *Empathize* (to understand the real concerns of stakeholders), *define* (to find out the deeper roots of the needs), *ideate* (to explore and generate solutions), *prototype* (to make tangible objects for some ideated solutions) and *test* (to evaluate the prototypes with the end-users and learn from them). The design-thinking stages may occur concurrently and are rapidly iterated, where the prototyped artifacts (e.g., products, services, tools or processes) are tested per iteration. The practical experiences gained with the prototyped artifacts in every design round inform the following round about how to improve the artifacts. Via improving the most viable concepts in more detail, the designer aims at attaining a viable product eventually.

The starting point in design-thinking is a good understanding of the practice field. Further, the design process is informed by the practice and is highly collaborative, where all stakeholders are involved in all phases of developing the artifacts. Involving all stakeholders, especially early on in the design process, prevents disappointments in that the artifacts do not cater stakeholders' real needs. This early involvement of stakeholders enables discarding suboptimal solutions as soon as possible. This so-called fail fast approach gears the design process towards producing viable products with high chance of user adoption. Giving users the opportunity to give their inputs, ideas and viewpoints during the design process, is important in developing social services where citizens' participation and acceptance are of outmost importance (like in case of smart cities).

5.2. Need for trade-offs

There are many trade-offs needed to be made in designing data governance instruments in a given smart city context. Every design should determine which combination of data quality measures and data privacy, security and ethics related measures should be in place. Often, these measures are adversely related. For example, improving the anonymity of personal data requires reducing the data quality (like the exact age of 24 years old is transformed to age range of 20-29 years old). Adversely, a high data quality may result in personal data breach, see for example the attack described in [37]. As the design space is complex and multidimensional, many designs can be created. For example, for the privacy-by-policy measures, every design determines how far data subjects can be put in control of

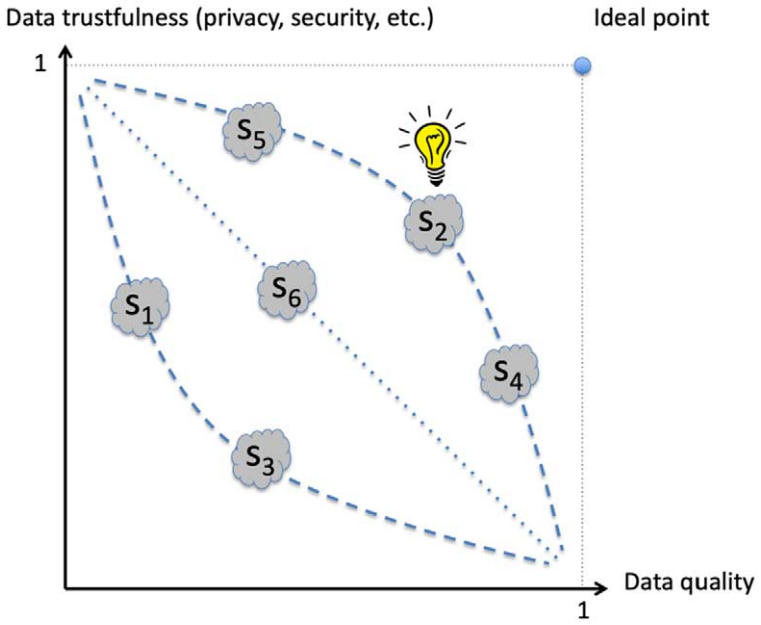

Fig. 5. An illustration of creating multiple design options, adopted with adaption from [7].

their data. This control can be at varying levels, ranging from full control to no control. Each choice, alone and in combination with other choices, has implications that should carefully be weighted and made.

Making tradeoffs among competing values is studied, for example, by [13] between secrecy and transparency and by [24] between privacy and public good. These works focus only on data collection technologies or on the control of the collected data. When the number of dimensions in the design space increases, as explained above, it might be difficult to come up with some design options and choose the most viable design in a deterministic way (i.e., only by, e.g., engineers and with engineering rules). Using design-thinking, we envision, one can integrate the true knowledge (i.e., the models and theories from science) with the how knowledge (e.g., the technological opportunities demonstrated by engineers) through an active process of ideating, iterating, and critiquing the potential solutions to make the right thing [54].

In Fig. 5, the idea of creating a series of artifacts, denoted by solutions $S_1, \ldots$ and S_6, is illustrated in a simple two-dimensional design space of data trustfulness versus data quality. In Section 5.1, we argued how design-thinking can help to derive these six viable design options by making trade-offs among various criteria and also how it can help the stakeholders to achieve consensus on a most viable one, for example, solution S_2 shown in Fig. 5.

5.3. Need for collaborations

So far, we have identified the importance of several building blocks for an appropriate data governance in a data ecosystem. These building blocks pertain to functionalities (Section 3) as well as organizational aspects (Section 4). Each of these building blocks entails several challenges and needs further elaboration for a specific data domain. The elaboration may evolve in several directions which are not necessarily divergent. As we expect that the involved parties (i.e., individuals and organizations) in a data ecosystem will be organized as a peer-to-peer organization and as argued there will be a shared responsibility among these parties, there should be full consensus among the member organizations (and the corresponding user groups such as citizens) about the direction along which the challenges should be elaborated and about the selected strategies to meet these challenges. This requires a sound collaboration among the entities within a data ecosystem. There are several means that might be deployed for an effective and efficient collaboration, such as proper communication, mutual understanding regarding the tasks and responsibilities, transparency, clear definition of tasks assignment, and exchange of best practices. The entities in a data ecosystem should come up with a set of mechanisms that they perceive as useful for their collaboration and

should prioritize these mechanisms. Furthermore, they should agree upon the deployment of each of the prioritized mechanisms.

With many stakeholders in a data ecosystem and different interests it is not always easy to come up with a set of mechanisms that earn the approval of all stakeholders. Design thinking, we suspect, can be exploited to obtain a prioritized list of the mechanisms approved by the stakeholders.

6. Conclusions

Today, a smart city is still mainly a collection of small-scale, independent pilot projects. This is far from sufficient to reach the goals of smart cities. Although individual issues, such as more charging stations, smart light masts, mobility hubs, and smart energy grids can be tackled in isolation, there is little insight in how these individual issues evolve when independent pilot projects are integrated, entailing for example additional issues like (data) ownership, data management, and data responsibility assignment across data chain parties. Therefore, we appealed for establishing an appropriate data governance within a data ecosystem. A data ecosystem can be regarded as the setting of a collaboration among all stakeholders, i.e., citizens and public as well as private partners within a data domain. Together they are responsible for the key functionalities of data governance in the data ecosystem. We elaborated upon provisioning the required data quality properties and enhancing the trust of stakeholders as the main data governance functionalities. These functionalities were chosen because their realization indicates another challenge for smart city ecosystems, namely: how to make trade-offs among contending values. Realizing these functionalities calls for, among others, setting up a suitable organizational structure for data governance, distributing the scarce data governance knowledge among organizations, and embedding the FAIR principle in a data ecosystem.

We touched upon the issues and solution directions that should be considered in meeting these challenges. For the realization of smart cities, particularly, it is necessary to make acceptable trade-offs between the data governance functionalities. To this end, there is a need for an efficient and effective collaboration among the stakeholders in the data ecosystem. We argued that design thinking can be a valuable method to achieve these goals.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments and suggestions on earlier versions of this paper.

Conflict of interest

None to report.

References

[1] V. Albino, U. Berardi and R.M. Dangelico, Smart cities: Definitions, dimensions, performance, and initiatives, *Journal of Urban Technology* **22**(1) (2015), 3–21. doi:10.1080/10630732.2014.942092.

[2] L. Anthopoulos, M. Janssen and V. Weerakkody, A unified smart city model (USCM) for smart city conceptualization and benchmarking, in: *E-Planning and Collaboration: Concepts, Methodologies, Tools, and Applications*, Vols 1–3, 2018, pp. 523–540, IGI Global. doi:10.4018/978-1-5225-5646-6.ch025.

[3] P. Assunção, A zero trust approach to network security, in: *Proceedings of the Digital Privacy and Security Conference*, 2019.

[4] J.C. Augusto, Smart cities: State of the art and future challenges, in: *Handbook of Smart Cities*, J.C. Augusto, ed., Springer, Cham, 2021. doi:10.1007/978-3-030-15145-4_95-1.

[5] M.S. Bargh, Realizing Secure and Privacy-Protecting Information Systems: Bridging the Gaps, Inauguration Lecture at Rotterdam University of Applied, Sciences (RUAS), The Netherlands, 2019, ISBN: 9789493012080, RUAS Press.

[6] M.S. Bargh and S. Choenni, On preserving privacy whilst integrating data in connected information systems, in: *Proceedings of the International Conference on Cloud Security Management (ICCSM'13)*, 2013, pp. 1–9.

[7] M.S. Bargh and S. Choenni, Towards applying design-thinking for designing privacy-protecting information systems, in: *Proceedings of the 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (IEEE TPS'19)*, Los Angeles, California, USA, 2019, pp. 12–14, (Co-located with IEEE CIC 2019 & IEEE CogMI 2019).
[8] M.S. Bargh, A. Latenko, S. van den Braak, M. Vink and R. Meijer, On statistical disclosure control technologies for protecting personal data in tabular data sets: A state-of-the-art study, *The Netherlands* (11 November 2020), Technical Report, reeks Cahier 2020-17: PU-Tools 2.0 project (nr. 3080) at Research and Documentation, Center, WODC and Hague, The, Available, https://www.wodc.nl/binaries/Cahier%202020-17_Volledige%20tekst_tcm28-470636.pdf.
[9] M.S. Bargh, R. Meijer and M. Vink, On statistical disclosure control technologies: For enabling personal data protection in open data settings, Technical Report, reeks Cahier 2018-20: PU-Tools project (nr. 2889), at Research and Documentation Center WODC, The, Hague, The Netherlands, 2017, Available.
[10] M.S. Bargh and P. Troxler, Digital transformations and their design: Renewal of the socio-technical approach, book chapter in Het Hoger Beroepsonderwijs in 2030: Toekomstverkenningen en Scenario's vanuit Hogeschool Rotterdam, Available, www.hr.nl/hbo2030.
[11] C. Batini, C. Cappiello, C. Francalanci and A. Maurino, Methodologies for data quality assessment and improvement, *ACM computing surveys (CSUR)* **41**(3) (2009), 16. doi:10.1145/1541880.1541883.
[12] P. Brous, M. Janssen and R. Krans, Data governance as success factor for data science, in: *Responsible Design, Implementation and Use of Information and Communication Technology. I3E 2020*, M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. Dwivedi and M. Mäntymäki, eds, Lecture Notes in Computer Science, Vol. 12066, Springer, Cham, 2020. doi:10.1007/978-3-030-44999-5_36.
[13] I. Büschel, R. Mehdi, A. Cammilleri, Y. Marzouki and B. Elger, Protecting human health and security in digital Europe: How to deal with the 'privacy paradox'?, *Science and Engineering Ethics* **20**(3) (2014), 639–658. doi:10.1007/s11948-013-9511-y.
[14] Z. Chen, L. Yan, Z. Lü, Y. Zhang, Y. Guo, W. Liu and J. Xuan, Research on zero-trust security protection technology of power IoT based on blockchain, *Journal of Physics: Conference Series* **1769**(1) (2021), 012039.
[15] S. Choenni, M.S. Bargh and N. Netten, Naar een datagedreven samenleving: uitdagingen, in: *Het Hoger Beroepsonderwijs in 2030: Toekomstverkenningen en Scenario's vanuit Hogeschool Rotterdam*, 2020, available at: www.hr.nl/hbo2030.
[16] S. Choenni, M.S. Bargh, N. Netten and S. Van Den Braak, Using data analytics results in practice: Challenges and solution directions, in: *Perspectives for Digital Social Innovation to Reshape the European Welfare Systems*, IOS Press, 2021, pp. 182–201.
[17] S. Choenni, M.S. Bargh, C. Roepan and R. Meijer, Privacy and security in data collection by citizens, in: *Smarter as the New Urban Agenda: A Comprehensive View of the 21st Century City*, J.R. Gil-Garcia, T.A. Pardo and T. Nam, eds, LNCS, Springer, 2016.
[18] S. Choenni, N. Netten and M.S. Bargh, Exploiting big data for smart government: facing the challenges, in: *Handbook of Smart Cities*, J.C. Augusto, ed., 2021.
[19] N. Chyi and Y. Panfil, A commons approach to smart city data governance: How Elinor Ostrom can make cities smarter, 2020, technical report from New America, Available, https://www.newamerica.org/future-land-housing/reports/can-elinor-ostrom-make-cities-smarter/.
[20] S. Cuno, L. Bruns, N. Tcholtchev, P. Lämmel and I. Schieferdecker, Data governance and sovereignty in urban data spaces based on standardized ICT reference architectures, *Data* **4**(1) (2019), 16. doi:10.3390/data4010016.
[21] D. Doneda and V.A. Almeida, Privacy governance in cyberspace, *IEEE Internet Computing* **19**(3) (2015), 50–53. doi:10.1109/MIC.2015.66.
[22] D. Eke and O.J. Ebohon, The role of data governance in the development of inclusive smart cities, in: *Societal Challenges in the Smart Society, Including Proceedings of International Conference on the Ethical and Social Impact of ICT*, M.A. Oliva, J.P. Borondo, K. Murata and A.L. Palma, eds, 2020, pp. 603–619.
[23] D. Everett, Weblog: Data Governance vs. Data Management: What's the Difference? 7 August 2019, Available, https://blogs.informatica.com/2019/08/07/data-governance-vs-data-management-whats-the-difference/.
[24] J. Fedorowicz, J.L. Gogan and M.J. Culnan, Barriers to interorganizational information sharing in E-government: A stakeholder analysis, *Information Society* **26**(5) (2010), 315–329. doi:10.1080/01972243.2010.511556.
[25] J. Gayness Clark, N. Lang Beebe, K. Williams and L. Shepherd, Security and privacy governance: Criteria for systems design, *Journal of Information Privacy and Security* **5**(4) (2009), 3–30. doi:10.1080/15536548.2009.10855873.
[26] M.R. Gibbs, G. Shanks and R. Lederman, Data quality, database fragmentation and information privacy, *Surveillance & Society* **3**(1) (2005).
[27] J.R. Gil-Garcia, T. Pardo and T. Nam (Eds.), *Smarter as the new urban agenda: A comprehensive view of the 21st century city*, Vol. 11, Springer, 2015.
[28] S. Gurses and J.V.J. van Hoboken, Privacy after the Agile Turn, 2 May 2017, Available. doi:10.31235/osf.io/9gy73.
[29] C. Harrison, B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J. Paraszcza and P. Williams, Foundations for smarter cities, *IBM Journal of Research and Development* **54**(4) (2010), 1–16. doi:10.1147/JRD.2010.2048257.
[30] ITU, Smart sustainable cities: An analysis of definitions. A, technical report by International Telecommunications, Union (ITU), 2014, available at: www.itu.int/en/ITU-T/focusgroups/ssc/Documents/Approved_Deliverables/TR-Definitions.docx.
[31] R. Kitchin, C. Coletta, L. Evans and L. Heaphy, Creating smart cities: Introduction, in: *Creating Smart Cities*, C. Coletta, L. Evans, L. Heaphy and R. Kitchin, eds, Routledge, London, 2018, pp. 1–18. doi:10.4324/9781351182409.
[32] O. Kwon, N. Lee and B. Shin, Data quality management, data usage experience and acquisition intention of big data analytics, *International Journal of Information Management* **34**(3) (2014), 387–394.
[33] Y. Kyratsis, R. Ahmad and A. Holmes, Technology adoption and implementation in organisations: Comparative case studies of 12 English NHS trusts, *BMJ Open* **2** (2012), e000872. doi:10.1136/bmjopen-2012-000872.
[34] L. Lupi, City data plan: The conceptualisation of a policy instrument for data governance in smart cities, *Urban Science* **3**(3) (2019), 91. doi:10.3390/urbansci3030091.
[35] J. Merino, I. Caballero, B. Rivas, M. Serrano and M. Piattini, A data quality in use model for big data, *Future Generation Computer Systems* **63** (2016), 123–130. doi:10.1016/j.future.2015.11.024.

[36] D. Mir, Designing for the privacy commons, in: *Governing Privacy in Knowledge Commons (Cambridge Studies on Governing Knowledge Commons*, M. Sanfilippo, B. Frischmann and K. Strandburg, eds, Cambridge University Press, Cambridge, 2021, pp. 245–267. doi:10.1017/9781108749978.011.
[37] A. Narayanan and V. Shmatikov, Robust de-anonymization of large sparse datasets, in: *Proceedings of IEEE Symposium on Security and Privacy*, IEEE, 2008, pp. 111–125.
[38] NCSC, Bereid u voor op Zero Trust, Factsheet FS-2021-02, versie 1.0, Dutch National Cyber Security Centre, 17 August 2021.
[39] H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford University Press, Stanford, CA, 2009.
[40] K. Nissim and A. Wood, Is privacy privacy?, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376** (2018), 6. doi:10.1098/rsta.2017.0358.
[41] J. Pan and Z. Yang, Cybersecurity challenges and opportunities in the new edge computing+ IoT world, in: *Proceedings of the ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization*, 2018, pp. 29–32.
[42] R. Price and G. Shanks, A semiotic information quality framework, in: *Proceedings of the International Conference on Decision Support Systems DSS04*, 2004, pp. 658–672.
[43] C. Raab, Revisiting the governance of privacy: Contemporary policy instruments in global perspective, in: *Regulation & Governance*, 2018, pp. 1–18. doi:10.1111/rego.12222.
[44] S. Rose, O. Borchert, S. Mitchell and S. Connelly, 2020, Zero trust architecture.
[45] M. Sanfilippo, B. Frischmann and K. Standburg, Privacy as commons: Case evaluation through the governing knowledge commons framework, *Journal of Information Policy* **8** (2018), 116–166. doi:10.5325/jinfopoli.8.2018.0116.
[46] I. Schieferdecker, N. Tcholtchev, P. Lämmel, R. Scholz and E. Lap, Towards an open data based ICT reference architecture for smart cities, in: *Proceedings of Conference for e-Democracy and Open Government (CeDEM)*, IEEE, 2017, pp. 184–193.
[47] G. Shanks and P. Darke, Understanding data quality in a data warehouse: A semiotic approach, in: *Conference on Information Quality*, University of Massachusetts Lowell, 1998, pp. 292–309.
[48] D.J. Solove, Understanding Privacy, 2008, Harvard University Press.
[49] O. Tene, Privacy: The new generations, *International Data Privacy Law* **1**(1) (2011). doi:10.1093/idpl/ipq003.
[50] A.M. Toli and N. Murtagh, The concept of sustainability in smart city definitions, *Frontiers in Built Environment* **6** (2020), 77. doi:10.3389/fbuil.2020.00077.
[51] Y. Wand and R.Y. Wang, Anchoring data quality dimensions in ontological foundations, *Communications of the ACM* **39**(11) (1996), 86–95. doi:10.1145/240455.240479.
[52] R.Y. Wang and D.M. Strong, Beyond accuracy: What data quality means to data consumers, *Journal of management information systems* **12**(4) (1996), 5–33. doi:10.1080/07421222.1996.11518099.
[53] M.E. Whitman and H.J. Mattord, *Principles of Information Security*, Cengage Learning, 2011.
[54] J. Zimmerman, J. Forlizzi and S. Evenson, Research through design as a method for interaction design research in HCI, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems – CHI*, Vol. 7, ACM, 2007, pp. 493–502. doi:10.1145/1240624.1240704.