A Literature Review of Data Governance and Its Applicability to Smart Cities

Yusuf Bozkurt
Reutlingen University
yusuf.bozkurt@reutlingen-university.de

Alexander Rossmann
Reutlingen University
alexander.rossmann@reutlingen-university.de

Zeeshan Pervez
University of the West of Scotland
zeeshan.pervez@uws.ac.uk

Abstract

Data governance have been relevant for companies for a long time. Yet, in the broad discussion on smart cities, research on data governance in particular is scant, even though data governance plays an essential role in an environment with multiple stakeholders, complex IT structures and heterogeneous processes. Indeed, not only can a city benefit from the existing body of knowledge on data governance, but it can also make the appropriate adjustments for its digital transformation. Therefore, this literature review aims to spark research on urban data governance by providing an initial perspective for future studies. It provides a comprehensive overview of data governance and the relevant facets embedded in this strand of research. Furthermore, it provides a fundamental basis for future research on the development of an urban data governance framework.

1. Introduction

The digitization of cities has been discussed for years within the concept of smart cities. Research on cities has developed rapidly in recent years, and the term 'smart city' is frequently used. The literature on smart cities reveals multiple perspectives and definitions of the concept. While some definitions focus on information and communication technologies (ICTs) as drivers and enablers of sustainable development, others focus more on socio-economic, governance, and multi-stakeholder aspects [1]. The technological discussion is mainly driven by the implementation of smart city services with the help of Internet of Things (IoT) technology and cloud computing [2, 3, 4, 5, 6, 7, 8]. As such, a silo perspective of the corresponding data is prevalent. A holistic view of the entire city might break down the nation of data silos in individual sectors' applications (e.g. smart solutions for waste management, traffic management and air pollution) and yield enormous potential in the area of data-driven decision-making, simulation and urban services. But, the baseline for such approaches, in parallel with the appropriate information and communication technology (ICT) infrastructure, are urban data. Urban data are highly heterogeneous in some aspects (e.g. format, source, quality, privacy requirements, quantity) [9, 10]. Thus, without proper management of data at the city or even national level, their potential cannot be fully realized, digitization projects will always prove challenging, and a holistic view of the entire city will not be realized. Data governance can help overcome these challenges and create semantic compatibility between different technologies and data silos and bring stakeholders together through standardized data processes, data models and policies [11, 12, 13]. While many publications deal with ICT architectures and IoT implementations in cities [9, 14, 15], only a few studies focus on data governance in an urban context [16, 17, 18]. We want to promote research on urban data governance by providing an initial perspective for future studies. Therefore, this literature review provides a comprehensive overview of data governance and the relevant facets embedded in this research stream. Consequently, this paper focuses on the following three research questions (RQ):

RQ1: How is the concept of data governance conceptually defined?
RQ2: How is current research structured with respect to data governance?
RQ3: Which dimensions of data governance research are applicable to smart cities?

We applied a machine-learning-supported systematic literature review method to answer these RQs. This approach facilitates analysing more than 600 articles and gaining insights from a large dataset that leads to answers to the RQs.

The rest of this paper proceeds as follows: Section 2 provides background information on smart cities and data governance. Section 3 positions the paper in relation to existing work, and section 4 describes the applied research method. Section 5 shows the results of the text-mining analysis and unpacks the big picture of the data governance research stream. Concentrating on data governance in the urban context, section 6 synthesizes the findings. Finally, the paper concludes

URI: https://hdl.handle.net/10125/79666
978-0-9981331-5-7

with a discussion and an outline of further research directions in section 7.

2. Background

As an umbrella concept, the term governance refers to the act of governing in public and the private sector [19]. It sets the framework for the self-regulation of a system. That includes processes, strategies and mechanisms that shape the balance of power, decision-making, conflict resolution and the management of resources [20]. In the city context, governance is more than an elected government making and implementing decisions about assets, strategies and policy instruments. It is about managing a complex network of public authorities, citizens, stakeholders, civil society and businesses, and achieving effective collaboration between these different actors to promote democratic decision-making, strengthen citizen participation, foster public dialogue and improve the quality of decision-making [21]. That includes the effective use of ICT and data for transparency, citizen participation and optimized processes. Some researchers view today's cities as data factories and the engine of the urban data ecosystem that continuously generates enormous data [9, 22]. These 'data factories' consist of various devices (e.g. IoT) and operational (legacy) systems of city infrastructure, city services and city administration. Urban data can be generated and gathered from a wide range of sources and can be distinguished by source, structure, format, time, size and access rights (closed, shared, and open). Thus, urban data can be specified as highly heterogeneous and extremely fast-growing [9, 10]. This situation leads to two connected topics: data management and data governance in cities.

A uniform and efficient data management framework is an integral component of smart cities. Some studies describe urban data management as consisting of the components (1) data acquisition, (2) data processing and (3) data publication [10, 14, 23, 24]. Additional data security and privacy and networking and computing technologies reflect cross-layer components [14]. These components contain various fields of action. For example, standards are necessary to enable consistent data acquisition across different data sources. Furthermore, data quality must be ensured, as poor data quality affects the entire data process. After raw data are gathered, they can be used by various applications and services. Therefore, providing a standardized (open) application programming interface (API) is essential to foster data reuse and avoid redundancy [9, 14].

While data management deals with technologies, systems and their implementation, data governance deals with strategic issues, such as required data management measures and responsibilities [25, 26]. Some definitions of data governance are available in scientific literature, but a common understanding is that data are a company asset, and data governance represents a strategic framework for roles, duties and systems [25, 26, 27, 28, 29]. From a practitioner perspective, data governance is a company-wide strategic framework for all data-relevant topics, in which guidelines, standards, processes, responsibilities and technologies represent the main action fields. Data governance and data management are closely interconnected, but they have a different focus. Data governance is the blueprint, while data management is the technical implementation of data governance [25, 30, 31, 32]. Drawing on IT governance theory, Khatri and Brown [26] define a data governance framework as consisting of five decision domains: (1) data principles, (2) data quality, (3) metadata, (4) data access and (5) data life cycle. The data principles domain derives appropriate standards, policies and guidelines that affect the actions of the other four domains. The International Data Management Association provides a highly comprehensive framework on data governance and data management, in which data governance lies at the centre of all data management activities and ensures balance and consistency [25].

3. Related work

The discussion on data governance is not entirely new, either in academia or in practice [25, 26, 28, 29, 30, 31, 32]. Therefore, we analysed relevant related studies (i.e. literature reviews on data governance) to better position our work in the body of knowledge and contribute to research by specifically addressing research gaps. For example, research has conducted literature reviews on data governance from a cloud-computing perspective [33, 34]. In their work, Saed et al. [34] examine data governance with a particular focus on security policies. However, they do not explain which databases, methodologies and search terms they used to conduct the literature review. Al-Ruithe et al. [33], also investigate data governance from a cloud-computing perspective. They describe in detail what criteria they used to search and select the articles. They focus on 52 highly cited articles. However, latest available papers were not included in the analysis. Further literature reviews on data governance are available in the health domain [35, 36]. Given that the health domain deals with personal and sensitive data, it is not surprising that research is already available there. Elliott et al. [35] describe state of the art in data governance policies for healthcare data warehouses. They explore a conceptual framework for data warehousing governance, with nine dimensions, in 15

identified papers. Holmes et al. [36] conducted similar work a year later. They examine 31 papers regarding data warehousing governance for clinical research in distributed research networks. The Data Governance Institute framework guides the analysis and provides the grounding framework. Lee et al. [37] conduct a literature review on data governance to identify governance factors for platform ecosystems. To this end, they examined 51 papers. Similar to Saed et al. [34], they do not mention the search criteria or methodologies. Alhassan et al. [38] conduct a comprehensive analysis of data governance by deriving the current state of research from six academic databases. The result is an overview of the frequency of data governance activities, represented by the dimensions of action, area of governance and decision domain. Unfortunately, it is not clear from the analysis which data governance activities are specifically identified. It is also unclear whether the selection of literature is limited to a specific sector. Nielsen [39] conducts a literature review on data governance to suggest areas for future development of data governance in the public sector. However, limitations of that work are that the derivation is based only on reading the abstracts of the 62 articles, resulting in limited insights.

Literature reviews are a significant scientific contribution because they provide a stable basis for scholars' own research and the research of others. They identify research gaps in the particular field and provide an agenda for further research [40, 41, 42, 43]. In contrast with the presented reviews, we differ in the methodology used, the number of papers analysed and the specific focus on smart cities. We contribute to research and practice by (1) providing an overview of the research structure of data governance, related theories and concepts; (2) providing an overview of the current fields of action of data governance in the smart city context; and (3) offering a starting point for future work in other fields with relevance to data governance.

4. Research method

In this paper, in addition to providing a detailed analysis of data governance in the smart city context, we aim to present a big picture of data governance research. Therefore, we developed a text-mining-supported systematic literature review based on prior work [42, 43]. Figure 1 illustrates the research process.

According to Kitchenham and Charters [43], a review process has three main steps. The first is planning the review, in which the motivation, the research questions and the search strategy are determined. The second step is to conduct the review. For this, we initially applied the search strategy shown in Table 1 in the Web of Science (WoS), ACM, IEEE, EBSCOhost and Emerald databases. We exported the results of each database directly as a RIS/BibTeX file, imported into our literature management software and cleaned of duplicates, which resulted in a final count of 612 articles.

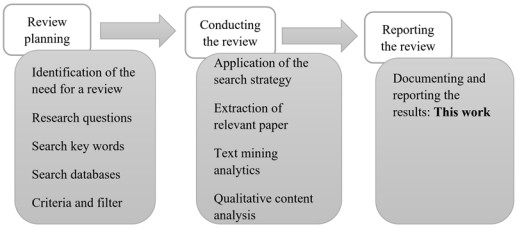

Figure 1. Research process

Table 1 reports the search strategy. In contrast to traditional literature review approaches [40, 42, 43], we integrated a text-mining phase in the research method. Through text mining, we were able to cluster 612 papers by thematic affiliation and generate insights from the entire dataset. Furthermore, a network analysis revealed the relationship of the topics to each other. After the text mining analysis, we identified smart city papers from the large dataset and conducted a full-text analysis of these relevant work using an inductive coding approach. Finally, we captured all results, which is the last step in the research process.

Table 1. Search strategy

In-/Exclusion criteria		**Search database**	
Language:	English	WoS:	436
Search string:	'Data Governance'	ACM	45
Time range:	To 2021	IEEE:	126
DB search fields:	Title, abstract, keywords	EBSCOhost:	138
Document type:	Journal article, conference paper	Emerald:	21
		Duplicates:	**154**
		Total:	**612**

5. Data governance research structure

We used RapidMiner Studio 9.9 for the implementation of the text-mining procedure. RapidMiner is a machine-learning and data-mining software that offers an extensive extension for text processing. It is freely available for research and educational purposes [44].

The hybrid text-mining-supported literature review begins with the structuring of the dataset to be analysed. The dataset should contain at least the title and abstract of the paper. Using any suitable software, the abstract texts are processed by tokenizing. This means that the abstract of each article is parsed into individual indexable elements. Subsequently, the dataset is cleaned by filtering out non-letters and stop words (e.g. 'the', 'this', 'me'). In addition, we filtered out generic words

such as ‘article’, ‘paper’, and ‘study’ to remove background noise from the dataset for cluster analysis. By tokenizing into individual elements, meaningful combinations such as ‘big data’ are lost. This also leads to individual words of the combined term appearing frequently and distorting the semantic comprehensibility, resulting in words such as ‘data’ appearing more frequently than others. To recover the semantic intelligibility, we form n-grams. For example, a bi-gram is formed by joining two consecutive words with an underscore (e.g. big_data). In our analysis, we formed tri-grams, which allowed us to achieve a high level of detail. Finally, the processed dataset can be analysed using data-mining methods such as K-means clustering to identify thematically related clusters. Therefore, in our case, we used the cosine similarity value as the clustering algorithm, which is particularly suitable for texts [45]. Finally, the cluster analysis output can be used to create a network graph based on each paper's similarity/distance value.

Figure 2 illustrates the result of the text-mining process as a word cloud. The most important keywords in the word cloud are data management, data quality, data sharing, decision making and data analytics. These are general topics without domain specification. Considering the keywords that make up a small part of the total, domains such as smart city, health data, cloud computing and big data governance can be identified. The word cloud is thus a first indicator of how research on data governance is structured.

Figure 2. Word cloud

Another artefact used to gain insights from the text-mining process is the clustering output, as shown in Table 2 with cluster ID/label, the number of papers per cluster and the words that make up the cluster. RapidMiner Studio 9.9 is able to determine the optimal number of clusters using the value ‘average within-centroid distance’. The higher this value, the more variance is contained in the clusters, indicating that the clusters are not thematically significant [46, 47]. From the ‘average within-centroid distance’ value and the interpretability of the words assigned to the cluster, we estimated the cluster number 10 for 612 articles as optimal through several iterations, taking into account human sensemaking to ensure robust results.

Table 2. Data governance cluster

Cluster ID/Label (No. of articles)	Significant terms
0/big data (120)	big_data; big_data_governance; data_analytics; big_data_analytics; data_analysis; data_governance_framework; social_media; data_processing; social_governance
1/data management (75)	data_management; data_sources; maturity_level; governance_processes; business_models; data_standards
2/data sharing (71)	data_sharing; data_integrity; access_control; machine_learning; data_access
3/operation and organization (22)	data_stewardship; customer_data; data_collection; cross_functional; higher_education_institutions; data_governance_practices;
4/master data (19)	master_data; master_data_management; data_management; data_quality
5/cloud computing (47)	cloud_computing; cloud_data; security_data; cloud_data_governance; data_security
6/decision making (92)	data_quality; decision_making; business_intelligence; information_technology; enterprise_data data_governance_model; reference_model; quality_control
7/corporate governance (31)	data_assets; information_governance; corporate_governance; governance_models; information_assets
8/health (49)	health_data; health_care; clinical_data; covid_pandemic; clinical_trials; web_based; health_data_governance
9/city and citizen (59)	data_protection; smart_city; personal_data; general_data_protection; data_protection_regulation; european_union

Cluster 0 forms the largest cluster, which contains topics on big data in combination with data analytics, data processing and data governance framework. Cluster 1 includes topics on data management, data sources and data standards. Cluster 2 is dominated by data sharing and data access topics and contains the keyword machine learning, which indicates that machine-learning applications are related to these topics. Cluster 3 lists the organizational and operational aspects of research with terms such as data stewardship,

data collection, and cross-functional and data governance practices. Cluster 4 represents the thematic of master data, management of master data and data quality. Cluster 5 represents the topical focus on cloud computing; an indication that cloud computing in a data security context plays a role in data governance is also visible through the keywords security data and data security. In cluster 6, keywords such as business intelligence, decision making and enterprise data are represented. Data governance thus seems to be intensely discussed in corporate decision-making (e.g. in the context of business intelligence activities). That the keyword data quality is represented indicates the importance of high-quality data for decision-making. Cluster 7 shows the strategic character of data governance. Data are considered an asset in the entire company, and corporate governance plays an important role [25, 26, 27, 28, 29]. Cluster 8 shows data governance in the health sector with the keywords health data, clinical data and health data governance. Finally, cluster 9 deals with data protection, personal data and data protection regulation. Here, the word 'smart city' also stands out, indicating that the data governance discussion in the smart city context centres on personal data and data protection. This is also shown in the literature review by Bozkurt et al. [48], where privacy and data protection are essential topics in the smart city context. Table 2 reflects how the discussion on data governance is structured in research and which thematic focal points are emerging.

The clusters can be represented in a network graph to find information about the connection of the thematic foci and the relationship to other clusters. For this purpose, we created a network graph in Figure 3.

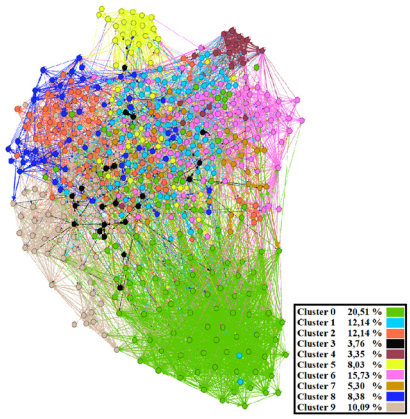

Figure 3. Data governance network graph

Cluster 0 is the largest cluster and thus forms the largest node. Although cluster 0 forms its own cluster, individual points are also scattered, indicating that topics in cluster 0, such as big data, data analytics, and data processing, are also discussed in other clusters. In the upper-right-hand corner, the topics of master data (cluster 4), data management (cluster 1) and data quality (cluster 6) are grouped. This indicates that these topics are also discussed together. Cluster 5 with cloud computing behaves similarly to cluster 0 with big data, in which there is a focus, but the yellow dots still appear distributed. Another consolation is formed by cluster 8 (health data) and cluster 2 (data sharing), which show a close interlocking, indicating that access control is an important issue, especially in the health sector. Furthermore, the proximity of cluster 9 shows a valuable insight that data protection and personal data also affect health data. Some researchers also understand smart health as a dimension in the context of smart cities [49, 50, 51], which justifies the proximity of cluster 8 to cluster 9.

6. Urban data governance

After presenting the broad research stream of data governance by building the big picture with a cluster and network analysis of more than 600 papers, this section aims to provide an in-depth content analysis for an urban context. To do this, we first screened the full text of all papers from cluster 9, which is strongly influenced by city-related topics. Among these 59 papers, we found 15 papers that proved relevant for our detailed analysis. According to Webster and Watson [42], a review is successful if it helps the reader understand the topic through the knowledge gathered. A concept-oriented structure helps highlight the main concepts in the literature, reveal patterns and identify research gaps [42]. An abstraction of individual sub-concepts aids in recognizing the currently discussed dimensions of urban data governance, identifying missing action fields, and contributing to the theory of data governance. For this purpose, we proceeded with an inductive coding approach [52, 53, 54] and recorded data governance concepts, ideas, action fields and problems, and concerns that emerge in individual papers by full-text reading. Incrementally, we grouped similar themes into broader categories, resulting in eight dimensions after several iterations. The coding process was carried out iteratively using an Excel file to record and categorise the individual themes. This synthesis relied primarily on expert knowledge drawn from relevant literature. Table 3 presents the results of the coding procedure.

A first dimension refers to policy, compliance and legal. Within this dimension, data ownership, data sharing, data rights and legal aspects are identified. A major part of the theme is data privacy, security and transparency of personal data. The privacy and transparent treatment of citizens' private data are considered and discussed in various contexts. For example, anonymization and re-identification of data are discussed by An et al. [55] and Calzada and Almirall

[56]. Privacy is often discussed in the literature as a concern in the urban context and as a requirement of governance [16, 17, 56, 57, 58]. A particular emphasis is put on sharing personal data between companies and institutions [55, 56, 57, 59, 60]. The General Data Protection Regulation (GDPR) is stated as an essential cornerstone [17, 55, 56, 61, 62, 63]. The necessity of well-defined data rights [55, 60, 62], well-defined agreements [55] and both visibility and audibility [56] are highlighted in this context. Finally, data ownership in the urban context is also addressed [17, 57, 61, 62]. Thus, the data sovereignty of data owners goes hand in hand [17, 58, 60, 61].

Another dimension that can be derived from the literature is the management of stakeholders in urban data governance. Research reveals the following stakeholders: municipal service providers [59, 60, 61], city government [55, 57, 59, 60, 61], private companies [55, 57, 59, 61], citizens and municipalities [55, 57, 59, 60, 61], and research institutions and non-profit organizations [55, 61]. In the context of smart cities, stakeholders take on an integral role. Different actors from different sectors have different goals and requirements. This interdisciplinarity also occurs in the field of data governance. Paskaleva et al. [17] describe the stakeholders within urban data governance as a cross-sectoral structure, making stakeholder collaboration a key challenge of urban data governance. Doneda and Belli [64] recommend a collaborative multi-stakeholder approach to urban data governance. In general, the collaboration of different stakeholders is considered necessary for a sustainable and successful implementation of a smart city [65, 66, 67]. Stakeholders are involved at different levels of the data process, from data collection to provision. The needs of the different stakeholders must be taken into account. To this end, it is important to know which stakeholders are active in which phase of the data processing and to ensure the necessary communication, regulations, cooperation guidelines and contracts. This leads to the question of the responsibilities and roles of the individual actors [61].

The dimension of ‘organization’ comprises three aspects. The first is the collaborative [59, 60] and comprehensive [55] character of data governance. The multidisciplinary stakeholders [64] in the organization must be considered, as described in the stakeholder context, as must the alignment [16] of stakeholders' needs and interests [55]. Data governance in organizations is also related to the transformation and change process [59]. The literature includes roles and responsibilities that can be attributed to the organization dimension. In particular, the following roles are mentioned in the literature: data steward, technology steward, governance officer, data committee, data provider [61], data consumer, data subject and data collector [55]. However, the literature neither explains nor describes the roles and responsibilities in detail. Thus, no firm conclusions can be drawn on the organizational structure of urban data governance.

Table 3. Urban data governance dimensions

Dimension	Key Content	Article
Policy, compliance and legal	Privacy, personal data protection, GDPR, transparency, security	[16, 17, 55, 57, 58, 59, 60, 61, 62, 63, 64, 68, 69]
	Data ethics, trust, justice	[55, 56, 58, 63, 64]
	Regulations, legal, data rights	[16, 17, 56, 58, 61, 63, 64]
	Sharing, ownership	[17, 56, 58, 59, 60, 61, 63, 64]
Stakeholder	Public service provider, citizen, authorities, research, companies	[16, 17, 60, 61, 64, 69]
Organization	Collaborative, comprehensive, alignment, multi-disciplinary, transformation, needs and interests	[16, 17, 55, 64, 70]
	Data steward, technology steward, governance officer, data committee, data provider, data consumer, data subject, data collector	[16, 57, 60, 61, 70]
	Responsibilities	[16, 56, 58, 61, 64, 69, 70]
Data classification	Big data, enterprise data, research data, gov. data	[16, 55, 61, 63, 68, 69]
Data quality	Data quality, misshape, asymmetry, inconsistency, misinterpretation, data reusability	[16, 55, 63, 69]
Data access	Open data, private data, internal data, public data, restricted data, open API	[16, 17, 55, 56, 58, 59, 61, 62, 64, 69]
Data management	Data life cycle, data generation, data collection, data storing	[17, 55, 58, 61, 63, 69]
	IoT, data model, architecture, data source, data monitoring, data linkage, data infrastructure	[17, 55, 56, 61, 62, 63, 69]
Principles	Standards, interoperability, integrability	[16, 55, 56, 57, 58, 61, 62, 64]

Urban data differ in many ways. Smart cities are considered as data factories and therefore show high heterogeneity. The literature shows the classification of urban data into big data [16, 55, 58, 64, 68], official institutional data, corporate data, research data [61], machine-generated data, human-generated data and organizational data [68]. Big data in the urban context is the most discussed. Lupi [16] lists a classification of small, medium-sized and big data. Small data is understood as data generated by people from visual and textual media (departmental project reports). Medium-sized data is understood as data from public administration, universities and cultural institutions with a social value published within transparency measurements and statistics.

Data quality can be identified as another dimension in data governance. In the literature, misinterpretation of data and false information due to poor data quality are discussed. The danger of manipulation of urban data is also deemed a challenge [55]. Data quality measures are necessary throughout the entire life cycle to achieve a consistent database for the city [16, 55, 63, 69].

Data access is a significant dimension because of the heterogeneity of urban data, whether it is the data source, the actors or the data structure and type. The topic can be divided into legal and technical perspectives. On the one hand, access rights of different stakeholders are discussed and named as a challenge in the urban context [55, 56, 60, 61, 63, 68]. This discussion is mainly considered in connection with data sharing and data exchange. On the other hand, technical issues such as the design of open APIs are also listed here [57, 58]. From the standpoint of data access, the following classifications are visible: private, internal, public, restricted data and open data. Especially open data are often mentioned [16, 57, 59, 60, 61].

Data management is the dimension that deals with the technical issues throughout the data life cycle. Data management is also considered the executive hand of data governance [57]. Data life-cycle issues encompassing data generation, acquisition, processing, storage and archiving can be identified from the literature. In turn, these are linked to broader themes such as data modelling, metadata management, data architectures and data infrastructures [17, 55, 56, 58, 61, 62, 63, 69].

To support these dimensions, principles such as establishing standards in all aspects of data handling [55, 56, 60, 61, 62, 63] and enabling high interoperability and integrability [16, 55, 60, 63] appear in the literature.

7. Discussion and future research directions

The goal of this literature review is to spark research on urban data governance by providing an initial perspective for further studies. Therefore, it provided a comprehensive overview of data governance and the relevant facets embedded in this research stream. Moreover, it provided a fundamental basis for future research on the development of urban data governance framework. With regard to RQ1, we first outlined the data-related challenges of cities and the need for data governance measures in cities in the background section. Subsequently, we conceptually defined the understanding of data governance by deriving it from the existing literature [25, 26, 27, 28, 29, 30, 31, 32]. By transferring existing knowledge on data governance from different fields to the urban context, a valuable contribution can be made to the development of urban data governance. In related work, we have reviewed existing literature reviews on data governance. We differ from existing papers on data governance in the methodology used, the number of papers analysed and the combined analysis of data governance in general and in an urban context. With a focus on RQ2 to develop the general research structure on data governance, a comprehensive collection of papers is required. Searching relevant academic databases resulted in the collection of more than 600 articles. This large number is difficult to process manually, so we added a text-mining phase to our literature review method. We clustered the articles by their thematic affiliation using K-means. The cluster analysis with 10 clusters thus shows the discussed areas of data governance in research and which topics are discussed frequently (Table 2). Furthermore, the network analysis shows which superordinate domains (clusters) are closely linked or have little connection (Figure 3). These artefacts provide insights into action fields and domains of data governance on a meta-level. The identified domains/clusters are valuable in searching for data governance practices and their transferability to the smart city sector. Cluster 9 contains relevant papers for a detailed analysis of urban data governance. Addressing RQ3, we identified 15 articles with a clear urban focus within cluster 9 and analysed them in a full-text reading process. We abstracted the individual topics found in these articles and categorized them into eight dimensions through inductive coding. The dimensions ‘policy’, ‘compliance’ and ‘legal’ appear in 14 of the 15 papers. Here, we observe a correlation with the smart city literature, in which data protection and privacy are also dealt with in different contexts [14, 48, 71]. The European GDPR is an essential element in data protection and privacy [55, 56,

61, 62, 63]. The open data movement of cities and the use of IoT in urban infrastructures (e.g. cameras for parking surveillance) raise questions about data sharing, data access and data ownership [61, 65, 72, 73]. That is not an easy challenge given the multi-stakeholder landscape in the urban context. Data governance implies a transformation process and is not a temporary project but a deeply rooted part of the organization [25, 29]. To meet the interests of multidisciplinary stakeholders sustainably, organizational structures with responsibilities, roles and communication levels are necessary in urban data governance. According to Batty et al. [73], smart governance is the required overarching intelligence that connects the other smart city dimensions. It aims to improve the quality of life by involving different stakeholders. To achieve this goal, ICT solutions are used to connect relevant stakeholders and make the city 'smarter'. However, applying ICT to existing urban infrastructures does not necessarily make a city smart [74]; instead, it is about how ICT uses data to drive growth and guide urban development processes [75]. Therefore, data governance plays an essential role in smart city research and practice.

The full-text analysis shows that data governance should support the interests of citizens and the sustainable development of cities. Unfortunately, the discussion provides little evidence on how data governance for cities actually look like and how it can be implemented. Except for Lupi [16] and Paskaleva et al. [17], the remaining papers explore only concerns, issues and challenges related to data and the role of data governance practices. Therefore, a framework that supports cities in implementing data governance would be a fruitful future research direction. The smart city literature focuses strongly on IoT applications, ICT reference architectures and smart city frameworks [3, 4, 14, 22, 66, 73], but a focus on urban data governance is missing. Thus, a possible research avenue is to develop a data governance framework for cities that enables integrated data management. In this sense, research could develop an adaptive reference model for urban data governance that considers the dimensions of Table 3 refined with existing data governance reference models from other domains and expert interviews. Developing models can be expensive and time-consuming for companies and institutions. Therefore, reference models are a proven method to increase the effectiveness and efficiency of the actual modelling and subsequent implementation. Reference models claim to be generally valid and serve as ready-made solution schemes for coping with practical problems [76, 77, 78].

By implementing holistic data governance, cities should be able to build an integrated urban data basis and merge data silos. Doing so would open the door for many smart city solutions, from services for citizens to data-driven decision-making at the policy level and the implementation of urban digital twins. Data governance and data management are relevant for companies. Indeed, a city can benefit from this wealth of knowledge and make the appropriate adjustments to its digital transformation. Combining the current understanding of smart cities, data management and data governance, we can define urban data governance as follows: Urban data governance assesses all data-related issues of a city from a holistic perspective. The main goal is to ensure sustainable urban development by managing data in the interest of citizens and promoting business and services.

The findings of this review contribute to research and practice by (RQ1) conceptualizing the construct of data governance, (RQ2) outlining the current structure of research on data governance, and (RQ3) providing a detailed analysis on the dimensions of urban data governance. The findings provide a basis for further research and practical implementation. The identified gaps and future research directions can help other researchers to focus their work effectively. In addition, practitioners can use the results to implement data governance in their projects and strategies. For example, when developing a smart city strategy, the results of this work can help address the need for action on data governance and take the dimensions of data governance into account. Beyond contributing to the knowledge base of data governance, our text-mining-supported literature review methodology can be adapted to other research areas. With text mining, we were able to present the entire collected corpus on data governance in a structured way and glean insights into the most frequently mentioned data governance topics, thematic structures, and thematic clusters' relationships.

8. References

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

[2] Kim, T., C. Ramos, and S. Mohammed, "Smart City and IoT", *Future Generation Computer Systems 76*, 2017, pp. 159–162.

[3] Gaur, A., B. Scotney, G. Parr, and S. McClean, "Smart City Architecture and its Applications Based on IoT", *Procedia Computer Science 52*, 2015, pp. 1089–1094.

[4] Sanchez, L., L. Muñoz, J.A. Galache, et al., "SmartSantander: IoT experimentation over a smart city testbed", *Computer Networks 61*, 2014, pp. 217–238.

[5] Mitton, N., S. Papavassiliou, A. Puliafito, and K.S. Trivedi, "Combining Cloud and sensors in a smart city environment", *EURASIP Journal on Wireless Communications and Networking*, 2012, pp. 247.

[6] Lea, R., and M. Blackstock, "City Hub: A Cloud-Based IoT Platform for Smart Cities", *IEEE 6th International Conference on Cloud Computing Technology and Science*, 2014, pp. 799–804.

[7] Petrolo, R., V. Loscrì, and N. Mitton, "Towards a smart city based on cloud of things, a survey on the smart city vision and paradigms", *Transactions on Emerging Telecommunications Technologies 28*(1), 2017.
[8] Taleb, T., S. Dutta, A. Ksentini, M. Iqbal, and H. Flinck, "Mobile Edge Computing Potential in Making Cities Smarter", *IEEE Communications Magazine 55*(3), 2017, pp. 38–43.
[9] Moustaka, V., A. Vakali, and L.G. Anthopoulos, "A Systematic Review for Smart City Data Analytics", *ACM Computing Surveys 51*(5), 2019, pp. 1–41.
[10] Liu, X., A. Heller, and P.S. Nielsen, "CITIESData: a smart city data management framework", *Knowledge and Information Systems 53*(3), 2017, pp. 699–722.
[11] Huang, S., G. Wang, Y. Yan, and X. Fang, "Blockchain-based data management for digital twin of product", *Journal of Manufacturing Systems 54*, 2020, pp. 361–371.
[12] Wache, H., and B. Dinter, "The Digital Twin – Birth of an Integrated System in the Digital Age", *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020, pp. 5452–5461.
[13] Singh, S., E. Shehab, N. Higgins, K. Fowler, T. Tomiyama, and C. Fowler, "Challenges of Digital Twin in High Value Manufacturing", *SAE Technical Papers*, 2018.
[14] Gharaibeh, A., M.A. Salahuddin, S.J. Hussini, et al., "Smart Cities: A Survey on Data Management, Security, and Enabling Technologies", *IEEE Communications Surveys & Tutorials 19*(4), 2017, pp. 2456–2501.
[15] Hashem, I.A.T., V. Chang, N.B. Anuar, et al., "The role of big data in smart city", *International Journal of Information Management 36*(5), 2016, pp. 748–758.
[16] Lupi, L., "City Data Plan: The Conceptualisation of a Policy Instrument for Data Governance in Smart Cities", *Urban Science 3*(3), 2019, pp. 91.
[17] Paskaleva, K., J. Evans, C. Martin, T. Linjordet, D. Yang, and A. Karvonen, "Data Governance in the Sustainable Smart City", *Informatics 4*(4), 2017, pp. 41.
[18] Eke, D.O., and O.J. Ebohon, "The Role of Data Governance in the Development of Inclusive Smart Cities", *Societal Challenges in the Smart Society*, 2020, pp. 603.
[19] Emerson, K., T. Nabatchi, and S. Balogh, "An integrative framework for collaborative governance", *Journal of Public Administration Research and Theory 22*(1), 2012, pp. 1–29.
[20] Weber, K., and C. Klingenberg, *Data Governance - Der Leitfaden für die Praxis*, Hanser Verlag, München, 2020.
[21] Bingham, L.B., T. Nabatchi, and R. O'Leary, "The new governance: Practices and processes for stakeholder and citizen participation in the work of government", *Public Administration Review 65*(5), 2005, pp. 547–558.
[22] Barns, S., "Smart cities and urban data platforms: Designing interfaces for smart governance", *City, Culture and Society 12*, 2018, pp. 5–12.
[23] Zheng, Y., L. Capra, O. Wolfson, and H. Yang, "Urban Computing", *ACM Transactions on Intelligent Systems and Technology 5*(3), 2014, pp. 1–55.
[24] Bellini, P., P. Nesi, M. Paolucci, and I. Zaza, "Smart city architecture for data ingestion and analytics: Processes and solutions", *IEEE 4th International Conference on Big Data Computing Service and Applications*, 2018, pp. 137–144.
[25] Henderson, D., and S. Earley, *DAMA-DMBOK: Data Management Body of Knowledge*, Technics Publications, Basking Ridge, New Jersey, 2017.
[26] Khatri, V., and C. V. Brown, "Designing data governance", *Communications of the ACM 53*(1), 2010, pp. 148–152.
[27] Cheong, L.K., and V. Chang, "The need for data governance: A case study", *ACIS 2007 Proceedings - 18th Australasian Conference on Information Systems*, 2007, pp. 999–1008.
[28] Hildebrand, K., *Daten- und Informationsqualität*, Springer Fachmedien, Wiesbaden, 2011.
[29] Weber, K., B. Otto, and H. Österle, "One Size Does Not Fit All - A Contingency Approach to Data Governance", *Journal of Data and Information Quality*, 2009.
[30] Data Republic, "Data Governance vs Data Management", 2019. www.datarepublic.com/resources/resources-guides/data-governance-vs-data-management
[31] Everett, D., "Data Governance vs. Data Management: What's the Difference?", 2019. www.informatica.com/blogs/data-governance-vs-data-management-whats-the-difference.html
[32] Kidd, C., "Data Management vs Data Governance: An Introduction", 2020. www.bmc.com/blogs/data-governance-data-management/
[33] Al-Ruithe, M., E. Benkhelifa, and K. Hameed, "A systematic literature review of data governance and cloud data governance", *Personal and Ubiquitous Computing 23*, 2019, pp. 839–859.
[34] Saed, K.A., N. Aziz, A.W. Ramadhani, and N. Hafizah Hassan, "Data Governance Cloud Security Assessment at Data Center", *4th International Conference on Computer and Information Sciences, ICCOINS 2018*, 2018, pp. 8–11.
[35] Elliott, T.E., J.H. Holmes, A.J. Davidson, P.-A. La Chance, A.F. Nelson, and J.F. Steiner, "Data Warehouse Governance Programs in Healthcare Settings: A Literature Review and a Call to Action", *eGEMs 1*(1), 2013, pp. 15.
[36] Holmes, J.H., T.E. Elliott, J.S. Brown, et al., "Clinical research data warehouse governance for distributed research networks in the USA: A systematic review of the literature", *Journal of the American Medical Informatics Association 21*(4), 2014, pp. 730–736.
[37] Lee, S.U., L. Zhu, and R. Jeffery, "Data Governance for Platform Ecosystems: Critical Factors and the State of Practice", *21th Pacific Asia Conference on Information Systems*, 2017.
[38] Alhassan, I., D. Sammon, and M. Daly, "Data governance activities: an analysis of the literature", *Journal of Decision Systems 25*, 2016, pp. 64–75.
[39] Nielsen, O.B., "A Comprehensive Review of Data Governance Literature", *Selected Papers of the IRIS*, 2017.
[40] Okoli, C., "A Guide to Conducting a Standalone Systematic Literature Review", *Communications of the Association for Information Systems 37*, 2015.
[41] Paré, G., M.-C. Trudel, M. Jaana, and S. Kitsiou, "Synthesizing information systems knowledge: A typology of literature reviews", *Information & Management 52*(2), 2015, pp. 183–199.
[42] Webster, J., and R.T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review.", *MIS Quarterly 26*(2), 2002, pp. xiii–xxiii.

[43] Kitchenham, B., and S. Charters, *Guidelines for performing Systematic Literature Reviews in Software Engineering*, 2007.
[44] RapidMiner, "RapidMiner Studio", 2021. www.rapidminer.com/products/studio/
[45] Li, B., and L. Han, "Distance Weighted Cosine Similarity Measure for Text Classification", In *Intelligent Data Engineering and Automated Learning – IDEAL*. Springer, Berlin, 2013.
[46] RapidMiner, "Cluster Distance Performance", 2021. https://docs.rapidminer.com/latest/studio/operators/
[47] Minitab, "Interpret all statistics and graphs for Cluster K-Means", 2019. https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-k-means/interpret-the-results/all-statistics-and-graphs
[48] Bozkurt, Y., R. Braun, A. Rossmann, and D. Hertweck, "Smart Cities in Research: Status-Quo and Future Research Directions", *IADIS INTERNATIONAL JOURNAL ON WWW/INTERNET 18*(1), 2020, pp. 121–138.
[49] Trencher, G., and A. Karvonen, "Stretching 'smart': advancing health and well-being through the smart city agenda", *Local Environment 24*(7), 2019, pp. 610–627.
[50] Solanas, A., C. Patsakis, M. Conti, et al., "Smart health: A context-aware health paradigm within smart cities", *IEEE Communications Magazine 52*(8), 2014, pp. 74–81.
[51] Ding, D., M. Conti, and A. Solanas, "A smart health application and its related privacy issues", *Smart City Security and Privacy Workshop (SCSP-W)*, 2016.
[52] Gioia, D.A., K.G. Corley, and A.L. Hamilton, "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology", *Organizational Research Methods 16*(1), 2013, pp. 15–31.
[53] Corbin, J.M., and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria", *Qualitative Sociology 13*(1), 1990, pp. 3–21.
[54] Chandra, Y., and L. Shang, "Inductive Coding", *Qualitative Research Using R: A Systematic Approach*, 2019, pp. 91–106.
[55] An, X., S. Sun, W. Bai, and H. Deng, "Data integration in the development of smart cities in China: Towards a digital continuity model", *11th International Conference on Cyber Warfare and Security, ICCWS*, 2016, pp. 13–20.
[56] Calzada, I., and E. Almirall, "Data Ecosystems for protecting European citizens' digital rights", *Transforming Government: People, Process and Policy 14*(2), 2020.
[57] Thompson, N., R. Ravindran, and S. Nicosia, "Government data does not mean data governance: Lessons learned from a public sector application audit", *Government Information Quarterly 32*(3), 2015, pp. 316–322.
[58] Austin, L.M., and D. Lie, "Safe sharing sites", *New York University Law Review 94*(4), 2019, pp. 581–623.
[59] Artyushina, A., "Is civic data governance the key to democratic smart cities? The role of the urban data trust in Sidewalk Toronto", *Telematics and Informatics 55*, 2020.
[60] Rayi, P.S., R.L. Bothra, S. Wallace, and M. Venkatesh, "Smart city investments: A rapid decision framework for public private partnerships", *International Conference on Unmanned Aircraft Systems, ICUAS*, 2019, pp. 569–574.
[61] Cuno, S., 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, pp. 1–24.
[62] Kazmi, A., M. Serrano, and A. Lenis, "Smart governance of heterogeneous internet of things for smart cities", *International Conference on Sensing Technology, ICST*, 2019, pp. 58–64.
[63] Popham, J., J. Lavoie, and N. Coomber, "Constructing a Public Narrative of Regulations for Big Data and Analytics: Results From a Community-Driven Discussion", *Social Science Computer Review 38*(1), 2020, pp. 75–90.
[64] Doneda, D.C.M., and L. Belli, "Municipal Data Governance: An Analysis of Brazilian and European Practices", *Revista de Direito da Cidade 12*, 2020, pp. 40–63.
[65] Angelidou, M., "Smart city policies: A spatial approach", *Cities 41*, 2014, pp. 3–11.
[66] Chourabi, H., T. Nam, S. Walker, et al., "Understanding smart cities: An integrative framework", *45th Hawaii International Conference on System Sciences*, 2012, pp. 2289–2297.
[67] Ruhlandt, R.W.S., "The governance of smart cities: A systematic literature review", *Cities 81*, 2018, pp. 1–23.
[68] Sarker, M.N.I., M.N. Khatun, G.M. Alam, and M.S. Islam, "Big Data Driven Smart City: Way to Smart City Governance", *International Conference on Computing and Information Technology*, 2020, pp. 1–8.
[69] Zulkarnain, N., R. Kosala, B. Ranti, and S.H. Supangkat, "Big Data Governance for Building A Smart Cities", *International Conference on ICT for Smart Society (ICISS)*, 2019, pp. 1–5.
[70] Yulfitri, A., "Modeling operational model of data governance in government: Case study: Government agency X in Jakarta", *International Conference on Information Technology Systems and Innovation*, 2016, pp. 1–5.
[71] Zhang, K., J. Ni, K. Yang, X. Liang, J. Ren, and X.S. Shen, "Security and Privacy in Smart City Applications: Challenges and Solutions", *IEEE Communications Magazine 55*(1), 2017, pp. 122–129.
[72] Bachtiar, A., Suhardi, and W. Muhamad, "Literature review of open government data", *International Conference on Information Technology Systems and Innovation*, 2020, pp. 329–334.
[73] Batty, M., K.W. Axhausen, F. Giannotti, et al., "Smart cities of the future", *European Physical Journal: Special Topics 214*(1), 2012, pp. 481–518.
[74] Hollands, R.G., "Critical interventions into the corporate smart city", *Cambridge Journal of Regions, Economy and Society 8*(1), 2015, pp. 61–77.
[75] Caragliu, A., C. del Bo, and P. Nijkamp, "Smart cities in Europe", *Journal of Urban Technology 18*(2), 2011, pp. 65.
[76] Fettke, P., and P. Loos, "Der Beitrag der Referenzmodellierung zum Business Engineering.", *HMD - Praxis Wirtschaftsinform.*, 2005.
[77] Schütte, R., *Grundsätze ordnungsmäßiger Referenzmodellierung*, Gabler, Wiesbaden, 1998.
[78] Rosemann, M., and R. Schütte, "Multiperspektivische Referenzmodellierung", In *Referenzmodellierung*. Physica, Heidelberg, 1999.