Modelling Variations in Data Lifecycles – demonstrated in a Smart City Context

Claudia Roessing

School of Business

Maynooth University, Ireland

Supervisor: Prof. Markus Helfert

Head of School: Prof. Joseph Coughlan

A thesis submitted in fulfilment of the requirements for the degree of *Doctor of Philosophy*

September 2023

This thesis is dedicated to my uncle Adolfo, my aunts Maria do Socorro and Teresinha, in loving memory.

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Claudia Roessing
September 2023

Acknowledgments

I would like to take this opportunity to express my deep gratitude and appreciation to the following individuals and institutions without whom this thesis would not have been possible.

I am profoundly thankful for the guidance, expertise, and unwavering support provided by my supervisor, Professor Markus Helfert. Your guidance has been invaluable throughout this research journey. Your patience, insightful feedback, and encouragement have shaped my work.

I extend my gratitude to Professor Adegboyega Ojo and Zohreh Pourzolfaghar for their valuable insights and constructive criticism in my progress reviews throughout the years. Your expertise and dedication have enriched the quality of this research.

I am extremely grateful to all members of the BIGroup, for their support especially Viviana, Priyanka, Gueltekin, Rehan, Marco, Ruth, Mansoor, and Gabriel. I would also want to thank Carol, Paul and Lee-Ann from the Innovation Value Institute.

To my family and friends, who have provided unwavering support, love, and understanding throughout my academic journey, I am deeply grateful.

I acknowledge the financial support provided by Lero, Adapt and IVI. This support enabled me to carry out the necessary research and experiments for this thesis.

I would like to thank Priyanka, Rehan and Gueltekin, who have provided stimulating discussions and a sense of community throughout my academic journey. It was honour to share my PhD journey with you.

Dublin, September 2023

Abstract

The complexity of data management is increasing, posing additional problems to organizations as they need to gain insights from a large amount of information at their disposal in order to make better decisions and add value to their services or products.

Data lifecycles models include the phases and activities that data must go through from creation to disposal. There are several benefits of using these models. They are used to reduce complexity of planning and handling data, to assist stakeholders in better understanding of the data available to them, and to comprehend how data is transformed into knowledge. Different data lifecycle models are necessary to process data and to meet different processing requirements and objectives. Despite the benefits of data lifecycle models, the modelling needs to be improved in order to reflect these necessary variations and provide pertinent information to the involved stakeholders, and support them in decision-making.

Data lifecycle models are frequently incorporated into enterprise models to allow expressing the alignment of data lifecycles with services and technical infrastructures. A common approach to model enterprises is Enterprise Architecture (EA), which is a conceptual blueprint that presents a holistic view of an organization's business processes and IT assets, as well as their relationships. Data lifecycle models can

be applied in several domains, for example research, semantic web, open data, and smart cities.

With the increasing complexity of data exchanges, understanding data lifecycles has become increasingly important in the last few years. For instance, this is particularly noticeable in smart cities, which are a dynamic ecosystem that faces many challenges in order to meet the needs of its citizens. These challenges include managing various stakeholders, gathering data from various sources, integrating heterogeneous data, and needing different processing times. According to the literature, smart cities can be viewed as urban enterprises with complex systems that must be integrated across multiple domains.

A data lifecycle is relevant to manage business data throughout its lifecycle while taking into account the constraints of business processes and allowing data requirements to be identified. Existing EAs for smart cities do not identify concepts for describing and modelling data lifecycle variations, which are required to process data.

This study suggests a metamodel for improving the modelling of data lifecycles. By designing the proposed metamodel, the elements of a data lifecycle, its variations, and design requirements are identified. Furthermore, the concepts and their relationships identified to model variations in data lifecycles are demonstrated and evaluated in two case studies in this thesis.

As a result, this study contributes to a better understanding of the data lifecycle variations and their requirements.

List of Publications

- Roessing, C. and Helfert, M. (2019). The need for Mapping Data Classification Standards - Illustrated in the context of FIPS 199 and BS 10010. In Proc. of 24th EURAS Annual Standardisation Conference (pp. 397-407). K. Jacobs and P. Morone (Eds.) DOI: 10.13140/RG.2.2.20985.49766
- Roessing, C. and Helfert, M. (2021). A Comparative Analysis of Smart Cities Frameworks based on Data Lifecycle Requirements. In Proceedings of the 10th International Conference on Smart Cities and Green ICT Systems, ISBN 978-989-758-512-8, ISSN 2184-4968, pages 212-219. DOI: 10.5220/0010479302120219
- Roessing, C. and Helfert, M. (2022). Identifying requirements to improve data lifecycle modeling. SMARTGREENS' 2021 Springer Publication (Accepted for Book chapter publication in "Communications in Computer and Information Science" (CCIS)).

Collaboration

- Development of the Data Classification Framework in the standard below:
  - International Organization for Standardization [ISO] (2021) ISO/IEC TS 38505-3 Information Technology – Governance of data – Part 3: Guidelines for data classification, ISO. Available at: https://www.iso.org/standard/56643.html

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADM	Architecture Development Method
AMME	Agile Modelling Method Engineering
API	Application Programming Interface
ARIS	Architecture of Integrated Information Systems
BASIS	Big Data Architect for Smart Cities
BOLD	Big and Open Linked Data
BPMN	Business Process Model and Notation
CCTV	Closed Circuit Television Systems
DODAF	The Department of Defence Architecture Framework
DSR	Design Science Research
DSRM	Design Science Research Methodology
EA	Enterprise Architecture
EU	European Union
FEAF	The Federal enterprise architecture framework
GDPR	General Data Protection Regulation
ICT	Information and Communication Technologies
IEC	International Electrotechnical Commission
IoT	Internet of Things
IP	Intellectual Property
IS	Information Systems
ISO	International Standards Organisation
KPI	Key Performance Indicator
MDE	Model-driven engineering
MEMO	Multi-Perspective Enterprise Modelling
MOF	Meta Objective Facility

NIST	United States National Institute of Standards and Technology
OECD	Organisation for Economic Co-operation and Development
OMG	Object Management Group
PCI	Personal Cardholder Information
PHI	Protected Health Information
PII	Personal Identifiable Information
RQ	Research Question
TOGAF	The Open Group Architecture Framework
TS	Technical Specification
UML	Unified Modelling Language

Chapter 1

1. Introduction

Nowadays, millions of data is produced daily, with official data showing that around 97 zettabytes are produced each day. Indeed, the forecast until 2025 is that 181 zettabytes will be created each day (Statista Research Department, 2022). Data is crucial in assisting organizations to make better decisions, as it allows them to understand customers, performance, and process improvements. As a result, it is paramount to treat data as an asset, with appropriate planning as to how it will be used, processed, and stored (Attard *et al.*, 2016). The adoption of a *data lifecycle* assists organizations in managing data and this data management tool contains all stages and activities that data has to go through, from its creation, processing, archival, and/or disposal (Arass, Tikito and Souissi, 2017). In addition, data lifecycles can assist data managers and other stakeholders in dealing with the deluge of data.

Due to technological evolutions, data management systems evolved too, from simple file systems, to modern industrial database management system servers, to big data management systems (Poltavtseva, 2019). However, despite advances in this area, the representation of data lifecycles is still being made at a high level, as this data management tool is still modelled from a high abstract point of view, representing an ideal situation, and showing data as unproblematic (Carlson, 2014; Pouchard, 2015; Cox and Tam, 2018, Shah et al. 2021; Weber and Kranzlmuller, 2019).

A smart city is a complex ecosystem, which uses data and information technologies in order to improve certain functions, such as transportation, education, health care, and public safety for its citizens. In other words, it is a city that tries to use its resources in an effective way to improve citizens' lives by overcoming challenges faced by urban centres (Albino, Berardi and Dangelico, 2015; Lim et al., 2018; Paskaleva et al., 2017).

In the 1990s, governments and researchers began to use the term Smart City to promote cities as innovative (Ramaprasad et al., 2017). However, the term's definition has evolved over time. First, according to Nam (2011) and Eger (2009), a smart city was defined as one that used information technology to handle a number of city services. Later, the concept encompassed sustainability, quality of life, and citizen services (Bifulco, 2016; Ahvenniemi, 2017). As the term has evolved, several authors have defined a "smart city" based on its objectives and initiatives, such as energy efficiency, security, environmental sustainability, and so on (Moura and de Abreu e Silva, 2019).

It is paramount for a city to process different types of data and meet different requirements in data processing to provide services to its citizens. For instance, some services process data that allow the identification of citizens, e.g. traffic management systems, while others, such as, air quality service, processes data that is not as sensitive. Therefore, to offer services in a smart city, different versions of data lifecycles are used, to take these particularities into account to process data.

The aim of this research is to improve the modelling of data lifecycles, modelling their variations within enterprise architectures, and demonstrate them in the smart city domain. To achieve this, this current research will propose a metamodel, to take into consideration variations that have been found in this framework; for instance: objectives, phases, activities, and types of data, just to name a few (Christopherson et al. 2020; Shah et al., 2021).

A smart city is an ideal domain to conduct this research due to challenges faced by cities in transforming data into knowledge and overcoming challenges faced by urban centres. In this multi-stakeholder system, services from different domains, offered to citizens, collect data from different sources with different formats that need to comply with regulations, and privacy and security requirements (Lněnička and Komárková, 2019; Gharaibeh et al., 2017).

As such, an improved form of the data lifecycle can assist decision-makers, in order to have a better understanding of the stages of a cycle, to improve a process, to assist data managers to understand the necessity for different models to process data, as well, as being used as a form of communication between them to show the impact of changes (Christopherson et al., 2020). Moreover, it can facilitate the information flow between different services and lead to transparency for end-users of the services, thus increasing trust.

1.1 Research Background

1.1.1 Data lifecycles

Models are used to reduce complexity and represent a specific domain, thus facilitating communication between the different stakeholders involved. A data lifecycle is a data management tool, which contains phases and activities that data needs to go through in order to obtain required results and, consequently, to guarantee the collection of data for a specific use. It also preparing data for relevant users, thus meeting the requirements for quality and security (Sinaeepourfard, et al. 2016).

There are several different data lifecycle models (see Section 2.2), with many developed for specific domains such as smart cities, whilst others have been developed as multi-specific, to be used in any domain (Shah et al. 2021; Cox and Tam 2018). The main purpose of this data management tool is to reduce the complexity of planning and for understanding data processing, dividing tasks into phases and activities and providing flexibility for organizations to define their phases and activities based on their needs (Shah et al. 2021).

Currently, data lifecycles have a high level of abstraction, thus they do not offer a realistic representation of data processing (Cox and Tam 2018; Shah et al. 2021), omitting details, as most of them are developed in a generic way, in order to be used in several domains (Christopherson et al. 2020).

The representation of these models requires an improvement in order to represent relevant and necessary details during data management and to assist users in interpreting them (Bork and Roelens, 2021). The identification of these details will assist stakeholders to have a holistic view of data processing, facilitating communication between them, and helping them with decision-making. An example of a data lifecycle is the Abstract Personal Data Lifecycle (APDL) is illustrated below in Figure 1.

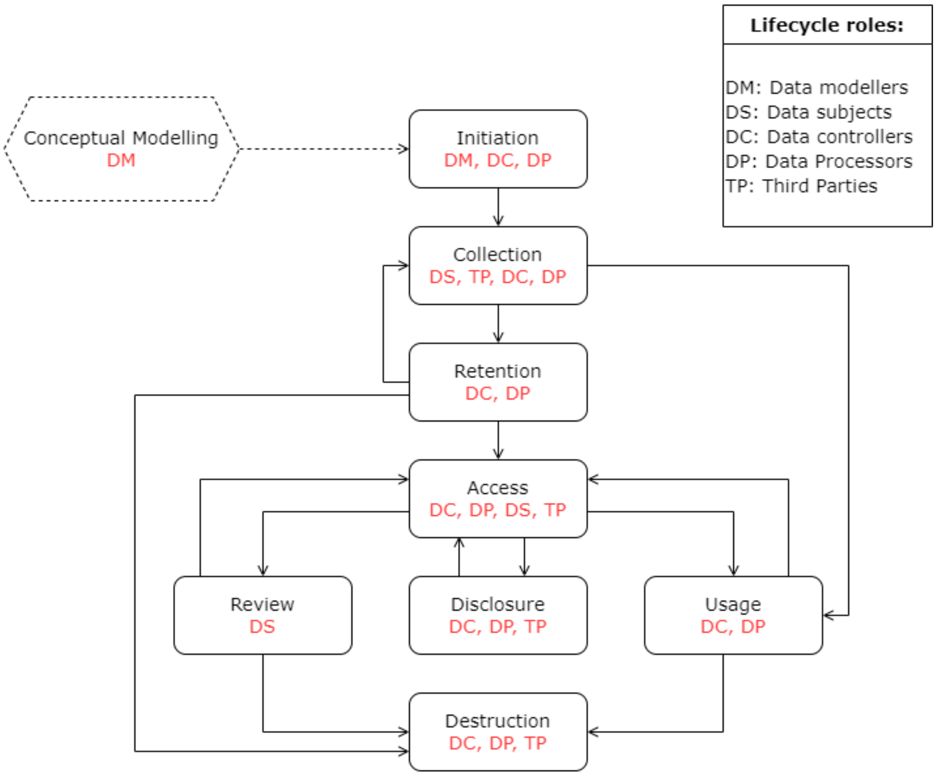

Figure 1 The APDLModel (Alshammari and Simpson, 2018)

1.1.2 Variation Driver

The concept of variability emerged years ago and is widely used in software product line engineering, as a solution to satisfy different stakeholder requirements (Pohl et al. 2005). It is related to differences existing in software or product that are taken into account during its development phase. This current research uses variability in order to model variations in data lifecycles and it focuses on the stages of identifying and modelling variability. Identifying variability in a model allows the representation of variations due to specific requirements. Pohl et al. (2005) identified four questions for the purpose of describing variability, what does vary, why does it vary, how does it vary, and for who is it documented?

- What does it vary? – variations points represent where variations occur in an artifact and are used to represent different implementations of features, processes, etc. (Rurua et al., 2019)
- Why does it vary? – Variations are due to certain factors and are identified as variation drivers. It is essential to understand the variation drivers in order to elicit variations.
- How does it vary? – Here, the options of the variations affected by the factors from the previous question are identified.
- For who is it documented? – Identification of the visibility of variability. Pohl et al. (2005) classify it as internal or external, depending on whether it is aimed at an internal or external stakeholder.

Taking into account the existing different types of data that must be processed in a smart city and different requirements that must be met, it is necessary to use different data lifecycles to meet the different needs of processing the data. The elements that influence the choice of phases and activities of a data lifecycle will be investigated in this study.

1.1.3 Enterprise Architecture

Enterprise architecture is a conceptual model which shows both the business and information technology, and their relationships in an organization. This tool is used to reduce organizational complexity and to assist communication between stakeholders (Lankhorst, 2017).

Nowadays, due to business competition, companies are searching for how to make decisions efficiently in a complex and dynamic environment, aiming to improve their processes, services, and products and, at the same time, to improve their decision-making (Barat et al., 2016). With the adoption of EA, organizations benefit in several ways, helping to highlight organization structure, business processes, and information systems (Lankhorst, 2017; Dumitriu, 2020), also capturing the relationship between information systems and business. By using EA, it is possible for an organization to identify its current stage and define a future stage of business, processes, and information systems infrastructure (Shanks, 2018).

In addition, EA assists in communication between stakeholders, as it maintains standardized processes, and assists in decision-making and in the digital transformation process (Korhonen, 2017). Currently, there are several frameworks that assist in the development of an EA (Urbaczewski, 2006; Dumitriu, 2020).

1.1.4 Enterprise Architecture for Smart Cities

A smart city is a conceptual city that uses information and communication technology in order to use resources efficiently, with the objective of minimizing problems of large urban centers and, consequently, increasing the quality of life of its citizens (Albino, 2015; Liu et al.,2017). However, management of a smart city brings with it several challenges, due to, processing of data from different sources, different formats, multi-stakeholders involved, services from different domains, and so on (Attard et al., 2016; Moustaka et al., 2018; Lim et al., 2018). In order to tackle these challenges, cities can be seen as enterprises and use EAs to reduce complexity and align their strategies with information and communication technology (Bastidas et al., 2017). Cities develop their EAs according to their needs and requirements, and they are presented in multilayers, with the purpose of describing processes, strategy, service, technology, information, etc.

1.1.5 Usage of data lifecycles in smart cities

Governments produce and collect a huge amount of data. With the deluge of data, it is essential that cities know how to create value in order to innovate and create services (Attard, 2016; Lim et al., 2018). However, literature and practice outline the challenges faced by cities regarding data management, due to the complexity and characteristics of this ecosystem (Gharaibeh et al, 2017; Liu et al., 2017).

Studies outline the benefits of using a data management tool to assist public and private organizations in data management (Shah, 2021). The main benefits are the division of the process into phases and activities, reducing complexity, and describing their sequences. A data lifecycle can assist stakeholders to plan what data to collect, how to collect it, how to use it, what data is available and storage plans, just to name a few (van Vestra, 2015; Carlson, 2014). Overall, the tool allows better management of data, thus meeting various requirements related to data management such as security and privacy (Alshammari and Simpson, 2018).

A smart city produces and collects data pertaining to multiple domains. The services offered by cities are designed in vertical silos separated by their domains. Thus, to process data, cities have to be in compliance with various regulations and other requirements (Hefnawy et al., 2017). Considering the benefits of using data lifecycles mentioned in section 1.1.1, along with the benefits promoted by EA, we can see that modelling data lifecycles in EAs in smart cities can assist these cities to better manage their data. It can thus facilitate the reuse of data, information flow between different services, and encourages the reuse of information. As a result, it will be possible to better integrate processes, people, services, and increase the interoperability of a smart city. Therefore, it is necessary for systems to not operate in silos in order to facilitate data flow (Hefnawy et al., 2017). Figure 2 illustrates data flow in the smart city domain.

Figure 2 Data Lifecycle Phases in the Smart City Domain

1.1.6 Usage of data lifecycles in enterprise architectures

A data lifecycle is relevant for managing data throughout its lifecycle while taking into consideration business process constraints and allowing data requirements to be determined in order to provide services or products to customers. Data lifecycle models are widely incorporated into enterprise models to represent how data lifecycles are aligned with services and technological infrastructures (Shah, 2021).

EAs are used by organizations for a variety of reasons including corporate strategic transformation, fostering business innovation, technology interoperability, compliance assessment, business-IT alignment, and technology standards management (Shanks et al., 2018).

Data lifecycles in EAs are critical for organisations to ensure that their data is aligned with their business goals, regulatory needs, and decision-making processes. Data lifecycle modelling in EAs enables organisations to adopt a consistent approach to managing their data assets. This will enable data processing to be aligned with organizations objectives/goals and will help decision-making (Demchenko et al., 2014).

1.2 Problem Statement and Motivation

Data lifecycle models are used to assist with data management understanding and planning. However, digital transformation has brought about a shift, from the traditional creation and delivery of services, to the use of digital technologies in order to provide better services. As such, modelling of data lifecycles needs to be improved, to consider current data processing requirements. For instance, in a smart city context, it is paramount to consider multi-domain services, diverse stakeholders, different data sensitivity, and so on.

However, despite advances in several areas, (for instance, data analysis), data modelling still has its limitations, and the way data have been modelled has not been adequately addressed in order to provide more information regarding data management. For instance, how is data collected, processed, reused, stored, and for how long? (Christopherson et al. 2020).

The aim of this current research is to address these concerns by improving the way data lifecycles are modelled, adding concepts that are relevant to data management. This research will investigate elements that are relevant for the representation of this model, its variations, showing their relationships and taking into consideration the data requirements of the smart city domain.

1.3 Hypothesis and Research Questions

Research questions are used to identify what a study is going to investigate (Thuan et al., 2019). To construct research questions to meet the objectives presented in Section 1.5, this current research followed the typology developed by Thuan et al. (2019).

The purpose of this research is to model variations of data lifecycles in EAs in the smart city domain. In order to achieve this, the hypothesis and research questions are presented below.

- **Hypothesis: Data lifecycle variations can be modelled in enterprise architectures to ensure the representation of different data processing requirements in the smart city context.**
- **Main Research Question:** *"How to support the representation of variations in data lifecycles in enterprise architectures in the context of smart cities?"*

Three sub-research questions are listed below with the aim of answering the main research question above.

- **Research Question 1:** What are the elements required to model a data lifecycle?

  The objective of research question one is to identify the elements necessary to model a data lifecycle using a literature review and practitioners' point of view to gather such data.
- **Research Question 2:** What are the concepts necessary to model variations in data lifecycles and how do we model them in enterprise architectures?

  Research question two focuses on results from research question one in order to develop a metamodel.
- **Research Question 3:** How do we evaluate the proposed concepts to support the modelling of data lifecycle variations in enterprise architectures?

  Research question three focuses on evaluating the metamodel concepts and their relationships found from research question two.

Figure 3 illustrates the research questions and content structure of this thesis, in order to develop an artifact and evaluate it.

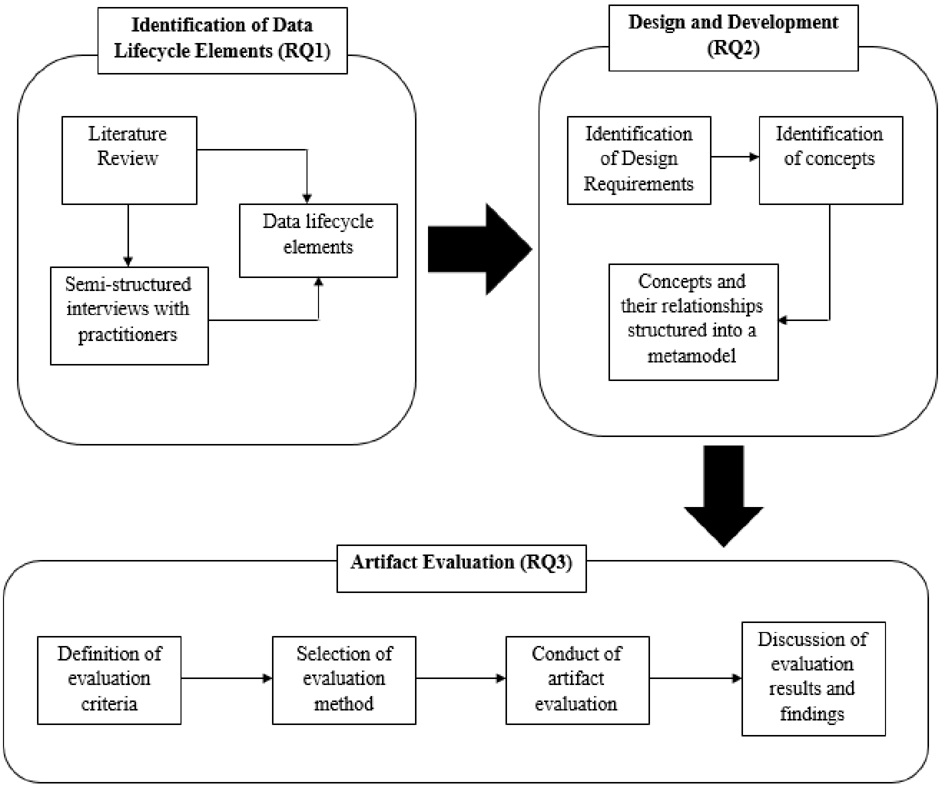

Figure 3 Research questions and content structure of this thesis

1.4 Research Challenges and research objectives

A smart city is a complex ecosystem, due to some unique characteristics, such as variety of types of data, its sensitivity, different types of processing, and the several stakeholders involved. As such, managing data in a smart city is a challenging task, especially in relation to the modelling of a data lifecycle. Therefore, the challenges identified in this current research are as follows:

Challenge 1: How do we define the design requirements to model variations of data lifecycles in enterprise architectures in the smart city domain?

Challenge 2: How do we capture the concepts and their relationships of smart cities required for data lifecycle modelling?

Challenge 3: How do we formally design the concepts and their relationships of smart cities to model a data lifecycle in enterprise architecture?

Challenge 4: How do we apply and evaluate the metamodel that supports modelling of variations of data lifecycles using real-world cases.

Research Objectives

The objectives of this research are based on the hypothesis and research questions presented in section 1.4 and are as follows:

- **Objective 1:**
  To identify design requirements for addressing data lifecycle modelling, taking into consideration data lifecycle elements and its variations.

- **Objective 2:**
  To define smart city concepts necessary to model a data lifecycle and the relationships between them.
- **Objective 3:**
  To design a metamodel that represents the concepts and relationships identified in objectives 1 and 2.
- **Objective 4:**
  To demonstrate usage of the metamodel and to evaluate it using real world use cases.

1.5 Thesis Contribution

This research designs a metamodel that enables modelling variations in data lifecycles in enterprise architectures, in the smart city domain. This section summarizes the contributions of this current research to academics and practitioners.

Academic Contribution:

- This study outlines the elements needed to model a data lifecycle and its variations identified in research and practice.
- This study enhances understanding that the choice of a data lifecycle is directly linked to the type of data processed and its requirements.
- It identifies concepts and their relationships necessary to model variations in data lifecycles in enterprise architecture.

Practice Contribution

- This research shows that identified concepts assist as a reference for the development of an EA.
- This research shows the relevance of identified concepts and their relationships for cities, to increase the understanding of information flow between different systems.
- The case studies show the importance of these concepts to assist relevant stakeholders to have a holistic view of data processed and thus help them in decision making and communication.
- It provides a taxonomy, which can help practitioners understand requirements for data management.

1.6 Research Gap

The purpose of data lifecycles is to provide information to those who make decisions, and to assist stakeholders in managing data (Shah et al. 2021).

The aim to leverage the modelling of data lifecycles in the smart city architectures is to take into consideration the novelty in the data management process in the smart city context and provide better modelling of the data flow in the smart city services as cities are becoming smart in order to improve citizen's lives using information and communication technologies (Oktaria et al. 2017; Hajjaji et al., 2021; Ahmad et al, 2022).

The importance of data for public and private organizations has amplified a demand for improvements in data lifecycle modelling, where it can provide more details to represent a specific situation or scenario. However, EAs do not identify the concepts (see Section 2.4) necessary to model such details relevant to data management, therefore these specific concepts should be added to data lifecycles, enabling cities to transform data into information and to create or innovate services for citizens.

1.7 Thesis Structure

The thesis consists of six chapters and is guided by the design science research methodology (DSRM). Figure 4 illustrates the sequence of chapters of this thesis and topics addressed by them.

The remaining chapters of the thesis are structured as follows: Chapter 2 provides a literature review to set the stage for this research. It analyses data life cycle modelling, identifying their limitations and drawbacks. The chapter also reviews data management challenges in the smart city domain and concludes with a review of enterprise architectures.

Chapter 3 presents the importance of using a research methodology in order to conduct research. The chapter presents different methodologies and the selection of a methodology to conduct this research based on its objectives and requirements.

Chapter 4 presents the design requirements extracted from the literature and concepts, in order to model them in EAs to answer RQ2.

Chapter 5 presents the application and evaluation of the artifact proposed in this research. The demonstration consists of the conduction of two case studies in city councils in Ireland. The chapter presents evaluation criteria used to evaluate the metamodel developed.

Chapter 6 concludes the thesis by summarizing its contributions and showing that this study has achieved its objectives. The chapter then presents limitations and future directions for this work.

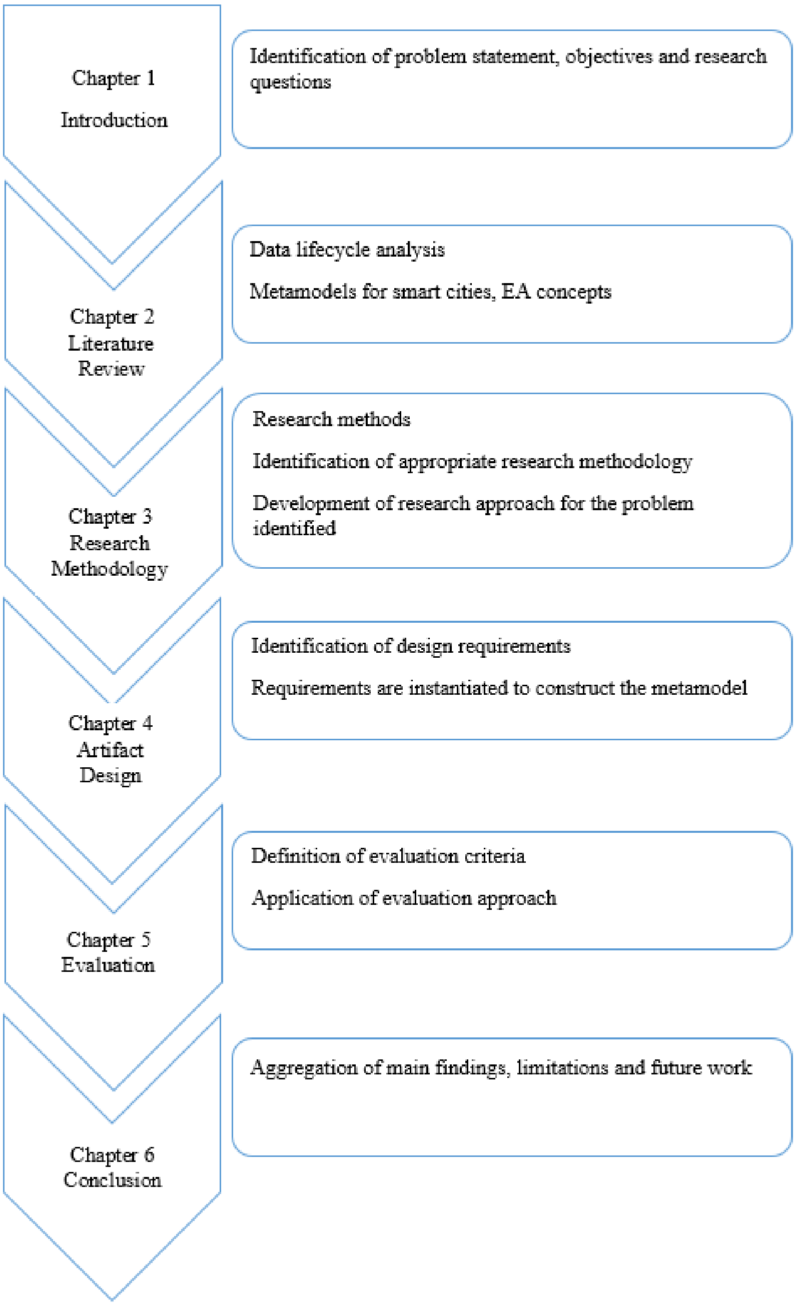

Figure 4 Thesis Structure

Chapter 2

2. Literature Review

In this chapter, we present a literature review, conducted with the purpose of meeting the objectives of this current research, as well as identifying problems. The review was conducted on data management in smart cities, data lifecycle modelling, and enterprise architecture for smart cities. More specific, the chapter will analyse current data lifecycles, as found in the literature, as well as analysing of different types of visual graphs of existing models and investigate the modelling limitations of these models. Subsequently, this thesis conducts a study to identify the modelling requirements of this model. The review also pays attention to challenges encountered in data processing in a smart city, and investigating EAs in the context of smart cities. Existing frameworks were analysed, to identify whether they offer concepts necessary to model a data lifecycle. This thesis focuses on the information layer, with the purpose of identifying a gap present in EAs for smart cities. Figure 5 illustrates the topics investigated during this review.

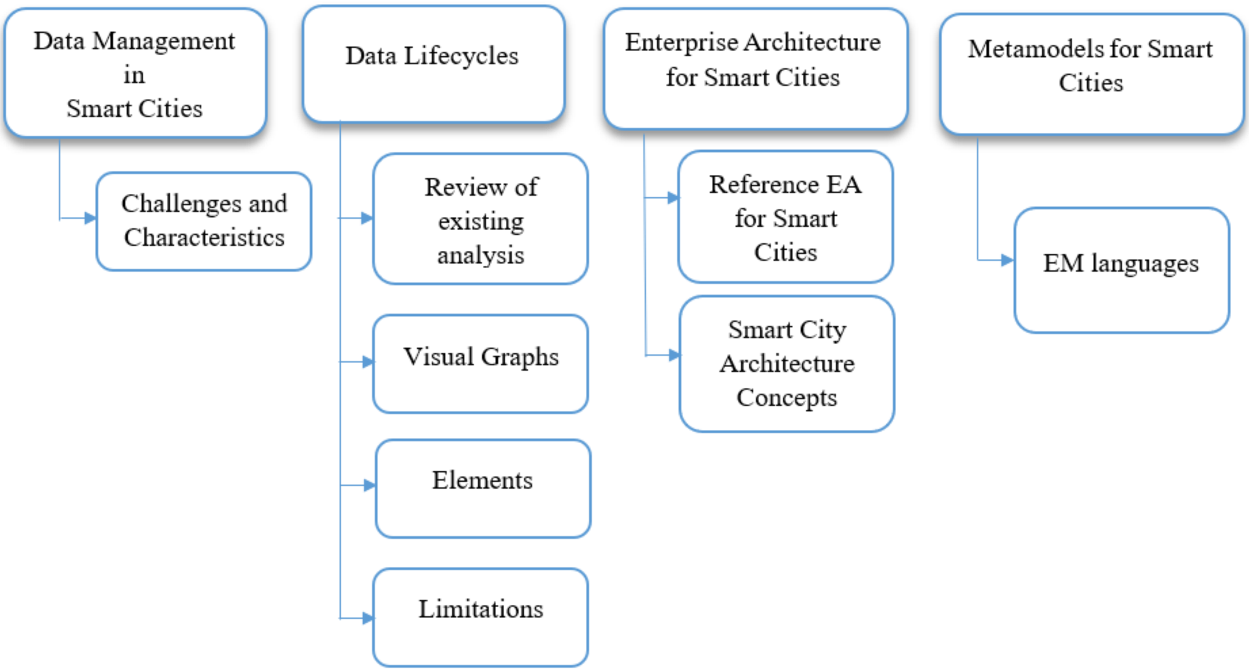

Figure 5 Structure of Literature Review

2.1 Literature Review Process

This section discusses the literature processes conducted by this research.

Data Lifecycle Modelling

The objective of this literature review was to explore the data lifecycle modelling. For this review, the researcher adopted a methodology proposed by Webster and Watson (2002) as illustrated on Figure 6. The first step was to define data sources (Springer Link, Google Scholar, Web of Science, Scopus) from where relevant studies were going to be collected. Posteriorly keywords were defined to be used as search strings in each library database provided. The keywords used were: data life cycle, data lifecycle, data framework, data management. Screening phase was conducted, where all duplicate articles were excluded, moreover, it was conducted a screening of abstracts from remained articles, in order to remove ones that were not relevant to this study. In the final step, 26 remaining articles are included for the review.

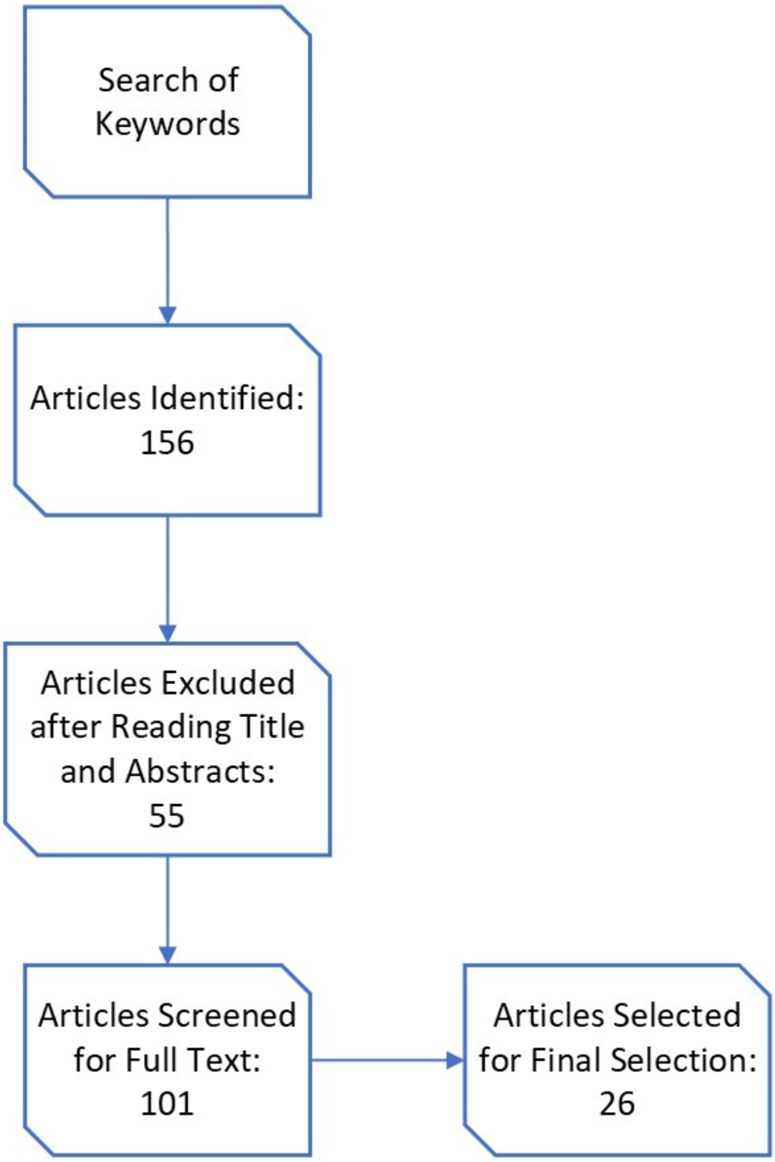

Figure 6 Literature Review – Article Selection

Design Requirements

The objective of this literature review was to identify the design requirements to model data lifecycles within enterprise architectures in the smart cities context. Figure 7 illustrates the steps of this article selection. Web of Science, Springer Link, Scopus and Google Scholar were the databases used for the search process. Articles not relevant to this study and repeated articles were excluded. Eighty nine articles were identified at the start of the review and thirty four were included in the review process.

Keywords used: data management, smart city, data management smart city, data processing, smart city services, smart city information architecture, smart city information system data processing regulations, data management architecture.

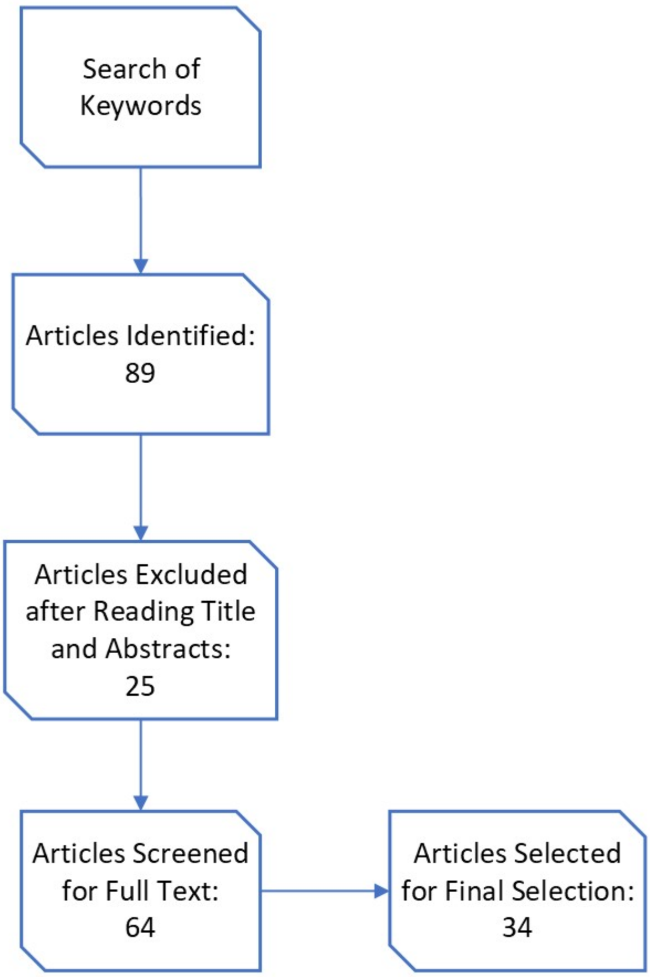

Figure 7 Design Requirements – Article Selection

Taxonomy

The objective of this literature review was to identify the elements to develop the taxonomy focussing on data management of smart cities as it can be seeing in Figure 8. Databases used were Springer Link, Scopus, Web of Science and Google Scholar. At the start of the process one hundred forty five articles were identified and forty nine articles were excluded based on the exclusion criteria and in the final step forty four articles were included in the review.

Keywords: data management, data processing smart city, data lifecycle smart city, data life cycle smart city, data processing regulations, data regulations.

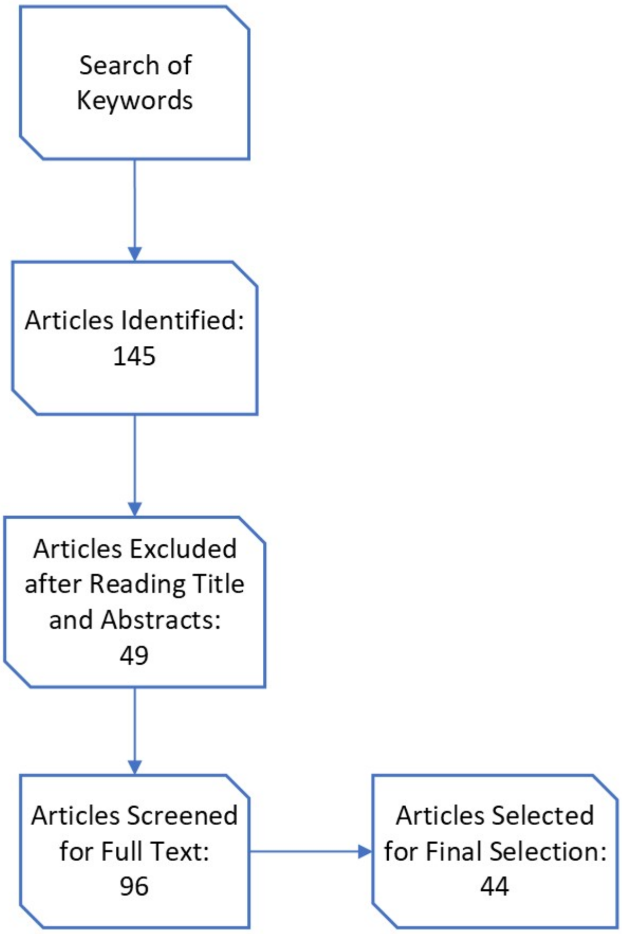

Figure 8 Taxonomy – Article Selection

2.2 Data management in the smart city context

Although the objective of smart cities is to use technology to improve citizens' lives by using resources wisely, advances in technology can bring about several challenges. One of the main challenges is data management (Gharaibeh et al., 2017; Chen et al., 2014; Hajjaji et al., 2021; Ahmad et al., 2022).

Big data and IoT make smart cities possible, however they also cause more difficulties in relation to managing data. As data is collected from various sources (sensors, smart meters, smartphones, CCTV cameras, etc.), this causes a complex problem when integrating and managing data (Siddiqa et al. 2016) (Chen et al. 2014) (Hajjaji et al., 2021) (Ahmad et al., 2022).

The utilization of big data in smart cities brings about more difficulties to manage data, due to its unique characteristics. It can be described by the 6 V's: volume, variety, velocity, variability, value, and veracity (Lněnička et al. 2017). In other words, big data is defined by the huge amount of data created from different types that are generated at high speed, therefore it is necessary to use tools and techniques to transform data into information, bringing value to services and products offered.

Other issues regarding data processing, include privacy, quality, and security of data (Gharaibeh et al. 2017; Attard et al., 2016; Ahmad et al., 2022; Paskaleva et al. 2017; Lim et al., 2018; Moustaka et al., 2018; Alshammari and Simpson, 2018). The following concerns have been raised, especially among the citizens (Table 1).

- What data is being collected?
- How is it being processed, especially personal data?
- Who has access?

- Is data being reused?
- For how long is data kept?
- What are the procedures to dispose of data, especially personal data?
- Is personal data being anonymized?
- Is General Data Protection Regulation (GDPR) or other relevant regulations being respected? (Voigt and Bussche, 2017)

Privacy and security are important issues to be considered in a smart city, however this concern is not restricted to city administrators, since citizens are also concerned about the privacy and security of their data collected and processed. Therefore, the use of data lifecycles is very relevant for data processing in smart cities and for promoting better knowledge of the information flow to offer services to citizens.

Table 1 Data processing concerns

References	Regulations Compliance	Transparency	Access of data	Privacy / Security	Disposal Procedures
Sarker, 2022		X		X	
Ahmad et al., 2022	X	X		X	
Cahlikova and Mabillard, 2019				X	
Lnenicka and Komarkova, 2019				X	
Sanchez-Corcuera et al., 2019		X		X	
Weber and Zarko, 2019				X	
Moustaka et al., 2018	X	X			X
Lim et al., 2018	X			X	
Demchenko et al., 2018		X	X	X	
British Standard, 2017		X	X	X	X
Shah et al., 2021	X	X			X
Sutherland and Cook, 2017		X	X		X
Charalabidis et al., 2018		X	X	X	
Alshammari and Simpson, 2018b	X	X	X	X	
Weber M., Žarko I. (2019)	X		X	X	
Hajjaji et al. (2021)				X	
Eryurek et al. (2021)	X	X	X	X	
Shahid et al. (2022)	X			X	X

2.3 Data Lifecycles

There are a variety of data lifecycles available for use in academia and practice. This data management tool has a key role in assisting stakeholders during a data management process. The variety occurs due to a need to meet different domains and different requirements in data processing (Shah et al. 2021; Christopherson et al. 2020).
Several studies have conducted analyses of this data management tool, with the main focus being to assist organizations in choosing the best model for their needs (Arass *et al.*, 2017, 2018; Cox and Tam, 2018).

2.3.1 A review of existing data lifecycles analysis

In this section, this study reviews existing analysis of data lifecycles available in the literature. As mentioned in section 2.2, there are several analyses of data lifecycles. However, most of these studies describe models, their applications, phases, limitations, and classification on certain criteria (Plale and Kouper, 2017; Sinaeepourfard, 2015; Pouchard, 2015; Möller, 2013; Arass et al., 2017; Elmekki et al., 2019; Cox and Tam, 2018).

Different types of analysis focus on different aspects. Möller (2013) analysed data life cycles from six different data-centric domains (multimedia, digital libraries, eLearning, knowledge and content management, ontology, and databases). Moreover, the author developed a new data lifecycle, the Abstract Data Lifecycle Model (ADM), which was used as a reference point, to compare models from data-centric domains. The study classifies models in the following characteristics: distinction data vs metadata, prescriptive vs descriptive, homogeneous vs heterogeneous, closed vs open, centralized vs distributed, lifecycle type, and granularity.

On the other hand, Carlson (2014) classifies models as individual, organization, or community, which are based on the type of audience that uses it.

Pouchard (2015) compared three data lifecycle models, with the aim of identifying gaps in data management, data curation, and data preparation phases. Elmekki et al. (2019) analysed the phases of five models, focusing on their contributions to political, economic, and social values.
Weber and Kranzlmüller (2019) analysed sixty-three models, with the aim of obtaining a common pattern to be used as quality indicators and to assess data lifecycles.
Cox and Tam (2018) merged four studies that focus on different aspects of data lifecycle models, purpose, visual graph, and element characteristics. To provide a critical analysis of nine data lifecycle models, the authors created a framework to compare models from different domains. The framework can be used to compare models on the following features: scope and point of view, elements and processes, and visualization. Fernández-López et al. (1997) analysed data lifecycles, focusing on how data changes along the process, and classified them into three categories: sequential, incremental, and evolving.

2.3.2 Different visual graphs of data lifecycles

Available studies show that only a minority analyse the structure and layout of data lifecycles and how they can affect a data management process. Model phases are connected in several ways, most presenting phases in a sequential way, where phases are arranged sequentially and subsequent phases can only start when a previous phase is completed. This means a new iteration of a cycle only occurs when all phases of a cycle are completed. It has been noted that some models have an incremental form, providing flexibility to data processing. It means, when data processing starts, it is not required for users to have full knowledge of the requirements, as this will happen in a gradual way.

It is possible to have evolving models, which means that iterations can happen at any time. Moreover, data can also change at any time during a cycle, and multiple iterations of a lifecycle can occur simultaneously.

To address the gap regarding the analysis of visual forms of data lifecycles, this current study conducted a literature review and found sixty-three data lifecycles from different domains, i.e. smart cities, big data, and research. The study considered models which contain the main phases identified in the Shah et al. (2021) study, which are collect, publish, and preserve. Subsequently, the models were used to identify different existing visual forms. Seven different visual forms were found and are described as follows:

- Circular – a model is circular when the last and first phases are connected, forming a circle and allowing the start of a new cycle.
- Non-circular – when a model contains a circular shape, though the first and last phases are not connected.
- Spiral – this model contains a spiral shape, indicating that information is incremented in each cycle.

- Integrated – an integrated model contains a combination of two or more different types of visual forms.
- Onion – this model contains different layers like an onion, and it is composed of activities, phases, stakeholders, and policies, just to name a few.
- Star – in this category, the model is shaped like a star, containing a central activity or data object in the middle surrounded by other activities or phases.
- Linear – here, a model has phases connected in a linear format, in the form of a straight line.

The list of identified data lifecycles and their respective classification are presented in Table 2.

Table 2 Data lifecycles classification

Data Lifecycle Model	**Circular**	**Spiral**	**Non circular**	**Integrated**	**Star**	**Onion**	**Linear**
A community-driven open data lifecycle model (Veenstra and Broek, 2015)	**x**						
ARMA International Life Cycle (CEOS, 2012)	**x**						
Big Data Analytics Lifecycle (Erl, Khattak and Buhler, 2015)							**x**
Big Data Lifecycle Management (Demchenko, De Laat and Membrey, 2014)							**x**
BLM Data Management Handbook (CEOS, 2012)	**x**						
Cassandra Ladino Hybrid Data Lifecycle Model (CEOS, 2012)	**x**						
COSA-DLC Model (Sinaeepourfard, Garcia, Masip-Bruin and Marín-Torder, 2016)							**x**
Data Framework (Defra, 2017)	**x**						
Data Lifecycle (Yu and Wen, 2010)							**x**
Data Lifecycle Management at Dell (CEOS, 2012)	**x**						
Data life cycle (Rüegg *et al.*, 2014)	**x**						
Data Lifecycle - UK Data Archive (Eynden, 2013)	**x**						
Data Life Cycle for Smart Cities (Emaldi *et al.*, 2015)							**x**
Data Security Lifecycle (CSA, 2011)	**x**						
Data Value Network (Attard, Orlandi and Auer, 2017)					**x**		
DataOne (Allard, 2012)	**x**						
DataTrain Model 1 (CEOS, 2012)							**x**
DataTrain Model 2 (CEOS, 2012)	**x**						
DDI Lifecycle (DDI Alliance)							**x**
Digital Curation Centre (DCC) Lifecycle Model (Higgins, 2008)				**x**		**x**	
e-research and data and information lifecycle (Wissik and Ďurčo, 2015)	**x**						
Enterprise Information Lifecycle	**x**						

Management (Chaki and Chaki, 2015)							
General Big Data LC (Khan *et al.*, 2014)	**x**						
Generic Science Data lifecycle (SDM, 2011)							**x**
Geospatial Data Lifecycle (OMB, 2010)	**x**						
Guide to Social Science Data Preparation and Archiving (ICPSR, 2005)							**x**
I2S2 Idealised Scientific Research Activity Lifecycle Model (Patel, 2011)				**x**			
IBM Lifecycle (IBM, 2013)	**x**						
Information pyramid (Creuseveau, 2015)							**x**
IWGDD's Digital Data Life Cycle Model (SDM, 2011)				**x**			
JISC Simon Hodson Research Data Lifecycle (Faundeen, John L., and Vivian B. Hutchison., 2017)	**x**						
John Faundeen & Ellyn Montgomery "Spins" (CEOS, 2012)		**x**					
Ku and Gil-Garcia (Sutherland and Cook, 2017)							**x**
Linear Data Life Cycle (SDM, 2011)							**x**
Lisa Dunn Data Flow Model (CEOS, 2012)							**x**
Manufacturing data lifecycle (Tao *et al.*, 2018)							**x**
Michigan State University Records Life Cycle Model (Faundeen, John L., and Vivian B. Hutchison., 2017)							**x**
NOAA Environmental Data Management Framework (Beaujardière, 2016)				**x**			
Open Government Data Life-Cycle (Attard *et al.*, 2015)	**x**						
Open Government Data Lifecycle (2017) (Belhiah and Bounabat, 2017)	**x**						
Personal Identifiable Information Lifecycle (Michota and Katsikas, 2015)	**x**						
Peter Fox Full Life Cycle of Data (CEOS, 2012)							**x**
Research Data Lifecycle by UK Data Service (UK Data Service)	**x**						
Research Data Lifecycle – University of Miami (University of Miami)	**x**						
Research lifecycle (RIN/NESTA 2010) (Cox and Tam, 2018)	**x**						

Research Lifecycle at University of Central Florida (University of Central Florida)	**x**						
SEAD Research Object (Plale and Kouper, 2017)	**x**						
Smart Data Lifecycle (Arass and Souissi, 2018)							**x**
Steve Tessler Data and System Lifecycle Models (CEOS, 2012)							**x**
The Abstract Data Life Cycle Model (Möller, 2013)				**x**			
The Big Data Life Cycle Model (Pouchard, 2015)	**x**						
The Data Lifecycle by ICSPR (ICPSR, 2012)			**x**				
The Data Life Cycle and Surrounding Data Ecosystem (Berman and Rutenbar, 2016)							**x**
The Digital Content Life Cycle (DigitalNZ)	**x**						
The integrated scientific life cycle of embedded networked sensor research (Pepe *et al.*, 2010)	**x**						
The Life cycle model of research knowledge creation (Humprey, 2006)	**x**						
The LOD2 Linked Open Data Life Cycle (Auer *et al.*, 2012)	**x**						
The Research360 Institutional Lifecycle Research Concept (Lyon, 2012)	**x**						
The scholarly knowledge cycle (Lyon *et al.*, 2004)				**x**			
The Research Lifecycle by University of Virginia Library(University of Virginia Library)							**x**
USGS Data Life Cycle Model (Faundeen, John L., and Vivian B. Hutchison., 2017)							**x**
USGS Data Management Plan Framework (DMPf) –(CEOS, 2012)	**x**						
Total	**32**	**1**	**1**	**6**	**1**	**1**	**21**

After analysing various models, this study identified a need to improve the modelling of data lifecycles, so that the model allows for the representation of the variations necessary to meet the different requirements in data processing. To accomplish this, it is necessary to understand the necessity of model variations and the reasons for this modelling requirement. As such, firstly it is necessary to understand different types of data, their classification and data processing requirements, in light of the fact that phases and activities of a data lifecycle change based on data management requirements.

2.3.3 Data Lifecycle Elements

In this section, the focus is to show the main elements of a data lifecycle that were identified during the literature review. They will be used to enhance the modelling of this data management tool. The literature review stated a lack of standardization in the modelling of this tool (Cox and Tam, 2018; Christopherson et al. 2020). Figure 9 and Table 3 depict the elements identified.

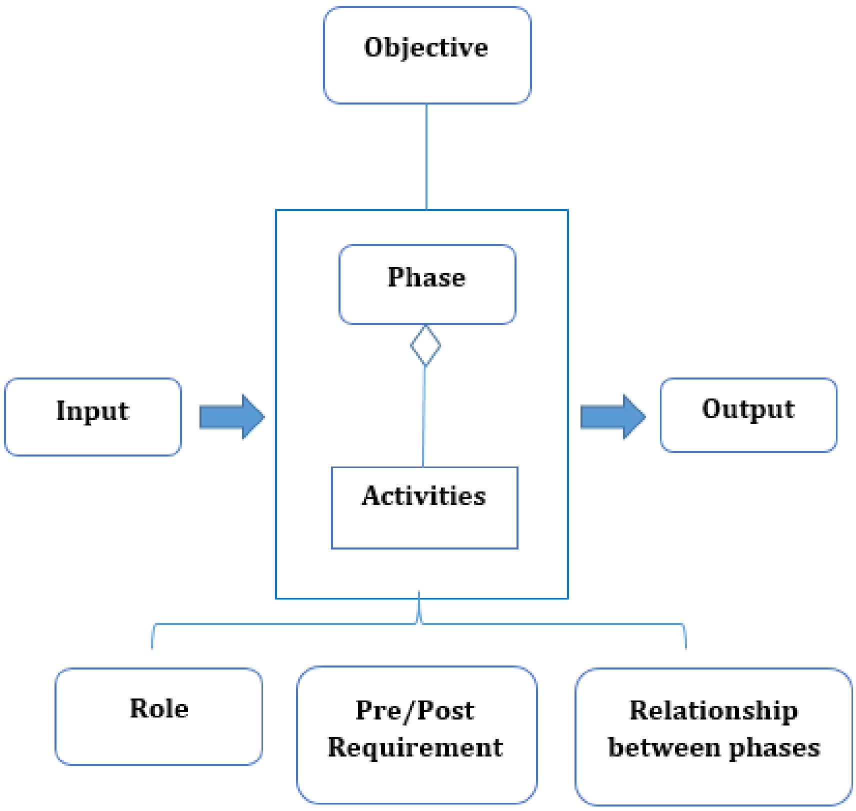

Figure 9 Data Lifecycle Elements

The main elements of a data lifecycle are listed as follows:

Objective – the purpose of a data life cycle, for what it is going to be used (Möller, 2013).

Phase – represents phases that data goes through in order to achieve a specific outcome (Attard et al., 2016).

Activities – are processes conducted during phases that prepare data to achieve a specific outcome (Attard et al., 2016).

Data input - the data that is fed into a phase or activity (van Veenstra and van den Brock, 2015).

Data output – is the data that was transformed after a phase or activity (van Veenstra and van den Brock, 2015).

Role - actor responsible for each activity or phase (Weber and Kranzlmuller, 2019).

Pre/Post-Requirement - it is used to determine whether a phase's activities have been completed successfully, in other words, if data has met a phase's objective, it can advance to the next phase, otherwise it must be processed again. These requirements are connected to the overall quality of each phase or activity (Alshammari and Simpson, 2018b).

Relationship between phases – informs the order of phases and activities which are necessary to process data to reach a final goal (Cox and Tam, 2018).

Other significant elements that must be considered in order to handle data were identified during the literature review. These elements are as follows:

Data Lifespan – related to the time at which data can be stored for usage (Charalabidis et al., 2018).

Regulations - the data regulations an organization has to consider in order to process data (Elmekki et al., 2019).

Category and sensitivity are related to data classification, which is the act of classifying data into categories; therefore, it can be protected more effectively based on data requirements (Digital Guardian; Carnegie Mellon University).

Data is considered the most valuable asset an organisation has, so it must be protected appropriately. For example protected health information (PHI), personal cardholder information (PCI), personally identifiable information (PII), and intellectual property (IP) are types of data protected by law and regulation. Failure to do so may have serious repercussions for an organization, including loss of customer trust, damage to the organization's reputation, and financial penalties (Digital Guardian; Carnegie Mellon University; British Standard, 2017; NIST, 2004; ISO, 2005; Cabinet Office, 2018a; Cabinet Office, 2018b; ISO, 2008).

During the literature review, it was also identified the elements that affect the choice of data lifecycle phases and activities, which will be represented in this study as a variation driver. Therefore, the variation driver is composed of important elements that must be taken into account when processing data, such as data type, regulations, lifespan, category, and sensitivity, as these elements influence the choice of data lifecycle phases and activities.

Table 3 Data Lifecycles Elements

References	Objective	Phases / Activities	Relationship between	Input / Output	Role	Pre / Post Requirement	Regulations	Category / Sensitivity	Quality	Data lifespan
Möller (2013)	X	X	X	X	X	X	X		X	X
Carlson (2014)	X	X	X	X	X					
Pouchard (2015)	X	X	X	X	X		X	X	X	X
van Veenstra and van den Brock (2015)		X		X	X		X	X	X	
Sinaeepourfard et al. (2016a)	X	X	X	X			X	X	X	X
Attard et al. (2016)	X	X	X	X	X		X		X	
Erl et al. (2016)		X	X	X					X	
Sinaeepourfard et al. (2016b)		X					X	X	X	X
El Arass et al. (2017)		X								X
Plale and Kouper (2017)		X			X				X	X
Liu et al. (2017)		X		X	X			X	X	
Faundeen et al. (2017)		X	X			X			X	X
Lnenicka and Komarkova (2019)	X	X		X	X		X	X		
Cox and Tam (2018)	X	X	X	X						
Khaloufi et al. (2018)	X	X	X	X	X	X		X	X	X
Alshammari and Simpson (2018a)	X	X	X	X	X	X	X	X		X
Alshammari and Simpson (2018b)	X	X	X	X	X	X	X	X	X	X
Charalabidis et al. (2018)	X	X	X	X	X	X	X	X	X	X
Mohamed et al. (2019)	X	X		X					X	
Department of Public Expenditure, NDP Delivery and Reform (2019)	X	X					X	X	X	X
Elmekki et al. (2019)		X	X	X			X		X	
Weber and Kranzlmuller (2019)	X	X	X		X					
Christopherson et al. (2020)	X	X		X	X	X			X	X
Shah et al. (2021)	X	X	X	X	X	X	X	X	X	X
Eryurek et al. (2021)	X	X	X	X	X	X	X	X	X	X
Sebastian-Coleman (2022)	X	X	X	X	X	X	X	X	X	X

2.3.4 Limitations of data lifecycle modelling

Several studies report the limitations of data lifecycles, stressing the need for models to show more information, which is necessary for data processing nowadays; for instance, to show information flow, and stakeholders involved (Christopherson et al. 2020, Shah et al., 2021). Furthermore, studies state a requirement to show transparency in data processing and how important it is to create value. It can be seen that most of the focus of research in relation to data management is directed to the analysis phase, looking for ways to improve development of new software or the emergence of new techniques (El Arass et al. 2017).

The improvement in the modelling of this data management tool will assist stakeholders to have a better understanding of information flow in a smart city scenario, assisting in communication and keeping objectives and services in this complex ecosystem aligned (Christopherson et al. 2020; Shah et al. 2021).

The literature review performed by this research shows the need to improve the modelling of this data management tool in order to address a gap in the data management field. The limitations of this tool are shown in Table 4. This research focuses on the lack of representation of this data management tool.

Table 4 Data lifecycles Limitations

References	Lack of flexibility to adapt	Do not consider data quality	Lack of reality representation	Do not consider data privacy/ security
Carlson (2014)	X		X	
Pouchard (2015)			X	
van Veenstra and van den Brock (2015)	X		X	
Sinaeepourfard et al. (2015)	X	X		
Bohli et al. (2015)				X
Sinaeepourfard et al. (2016b)	X			
El Arass et al. (2017)	X			
Faundeen et al. (2017)	X			
Plale and Kouper (2017)			X	
Paskaleva et al. (2017)				X
Liu et al. (2017)				X
Cox et al. (2018)			X	
Alshammari and Simpson (2018a)			X	X
Alshammari and Simpson (2018b)	X		X	X
Charalabidis et al. (2018)			X	X
Khaloufi et al. (2018)				X
Elmekki et al. (2019)		X		
Hou et al. (2019)				X
Christopherson et al. (2020)	X		X	
Gessa and Sancha (2020)			X	
Stoyanovic et al. (2020)		X	X	
Bossaller and Million (2023)			X	

2.4 Enterprise Architectures for Smart Cities

In this section, this current research examines existing architectures, frameworks, and layers of enterprise architectures for smart cities, in order to determine whether they support data lifecycle modelling and its variations.

The Zachman framework (Zachman, 1987) was one of the first frameworks developed for an information system architecture. Zachman stated that it is possible to have different types of descriptions of a same product; in other words, a same product can be described using different perspectives. In order to achieve that the framework has six perspectives or views, namely planner, owner, designer, builder, subcontractor, and user. And it also addresses six basic questions: what, how, where, who, and why.

The Department of Defence Architecture Framework (DODAF) (Chief Information Officer, 2010) is an architecture framework for the United States Department of Defence. Using this framework, stakeholders can focus on specific areas of interest in an organization. The main goal of this framework is to present information to several stakeholders in an understandable way, through viewpoints. The framework has eight views and it also provides a six-step architecture development process.

The Federal enterprise architecture framework (FEAF) (Chief Information Officers Council, 1999) is a US reference EA to facilitate the sharing of processes and information between Federal agencies and other government agencies. This framework is divided into business data, application, and technology architectures. It provides guidance to Federal agencies to structure their EAs and also allows flexibility for them to use their methods, tools, and work products.

The Open Group Framework (TOGAF) (The Open Group, 2018) is one of the most utilised frameworks for EA, and it can be used by any type of organization from small to large businesses including governments and non-government organizations. The framework provides guidance to develop, use, and maintain an EA. It uses an iterative process, allowing the usage of best practices and re-use of existing architectural assets. The central part of the framework is the Architecture Development Method (ADM), which provides architect's guidance regarding architecture development phases and ensures that a set of requirements are met.

TOGAF supports the development of four-layer architectures:

- Business architecture - where business strategy, governance, organization, and key business processes are described.
- Data architecture - describes logical and physical data assets of an organization and its data management resources.
- Application architecture - shows applications developed, and their relationship to the business processes of an organization.
- Technology architecture - describes IT infrastructure necessary to support business, data, and application services of an organization.

To summarise, the four major EA frameworks take different approaches and provide varying levels of detail. According to Bankauskaite's (2019) analysis, the frameworks received a low domain support score. This criterion refers to the use of EA in various domains.

Reference Enterprise Architectures for Smart Cities

The Espresso Smart City Reference Architecture (Cox et al., 2016) was created by a project co-funded by the Horizon 2020 Framework Program of the European Union. The project aims to define a smart city reference architecture for cities and communities, describing key elements in alignment with EU initiatives and worldwide standards. The framework adopts TOGAF framework for its development. The reference architecture contains nine layers: positioning services, sensing services, data services, application, business, consumers, and three cross-cutting services layers, namely security, technology, and supporting services. These layers include capabilities that are required by horizontal layers. The main purpose of the framework is to provide guidance to relevant stakeholders on how to use a reference architecture.

Cisco (Falconer and Mitchell, 2012) has developed the smart city framework for the purpose of encouraging public and private sector actions to implement smart city initiatives. The framework details the process for stakeholders to have an understanding of how cities operate in order to define city objectives and stakeholder roles and to understand how Information and Communication Technologies (ICT) can operate with city assets.

The framework is divided into four layers. Layer one at the bottom contains city objectives, which focus on improving social, environmental, and economic pillars. Layer two contains city indicators that are linked to city objectives. Layer three shows city components that detail city assets. Layer four defines city content, which maps objectives defined in layer one with best practices and policies that will be used to achieve identified objectives. The framework has a circular flow of information and its main benefits are definition of stakeholder roles and a detailed view of how cities work.

Costa and Santos (2016) developed a Big Data Architect for Smart Cities (BASIS), which aims to fill a gap in the technological layer and create multiple abstraction layers. The architecture contains three abstraction layers. The first layer is called conceptual layer, which presents several components that are important and necessary to extract Big Data, analyse it, and its availability. The technological layer describes technologies used for storage, analysis, and processing of data.

The infrastructural layer describes the physical part of the architecture; in other words, the hardware. The architecture also shows key roles, their relationships, cooperation to extract, process, and analyse big data.

Big and Open Linked Data (BOLD) (Lnenicka et al., 2017) was created to link and combine data, with the aim of gaining valuable insights from a vast amount of data. The authors define the most important requirements and their relationships present in a smart city ecosystem. Thirteen requirements are identified from a survey of architectures. The architecture is divided into four architectures. Business architecture describes E-government and governance architecture and open government processes. Application architecture describes application services. Data architecture contains programming models, data, Application programming interface (API), and data storage.

Technology describes smart ICT infrastructure and smart environment. Security and privacy + interoperability + evaluation and monitoring is a cross-cutting layer. The authors also suggest the use of ADM from TOGAF in order to develop architecture.

The frameworks that are discussed have advantages, but there are also some drawbacks.

BOLD presents a list of data lifecycle phases and activities but does not depict the connections between concepts. Espresso displays a list of data lifecycle phases but no relationships between concepts. BASIS depicts data lifecycle phases and identifies a variety of data, however there is no description of concepts or their relationships. Finally, although Cisco provides a description of the components of each layer, it does not show how they are linked to one another.

2.5 Smart City Architectures Concepts

As described in section 1.1.2, enterprise architectures are divided into layers, ranging from 3 to multi-layered architectures. The number of layers in architecture is defined according to the needs of the entities (Anthopoulos and Fitsilis, 2014).

Each layer illustrates different perspectives relevant to specific stakeholders (Helfert et al., 2018). Concepts are used to represent these different perspectives. TOGAF, which is used in this current research, divides an EA into four domains: business, data, information and technology.

This research focuses on the data and information layers, therefore this section will select concepts from existing EAs for smart cities, for the purpose of identifying missing concepts necessary to model variations in data lifecycles.

2.5.1 Business Concepts

This section selects concepts from existing EAs for smart cities to describe the main characteristics of city services presented in the business architecture layer.

Business unit – A business unit is a government organisation unit that provides services to citizens, such as a city council department that monitors air pollution (Cox et al., 2016; Barrutia et al., 2022).

City goal – The major objective of a smart city is to improve peoples' quality of life. To accomplish this, each subsystem has a purpose, such as reducing pollution, carbon footprint, energy usage, and so on (Attaran et al., 2022; Chen, 2023; Guimarães et al., 2020).

City service – a city service is a service provided to citizens by the city government or private organizations. Smart parking, bus transportation, health services, and smart metres are some examples of services (Sánchez-Corcuera et al., 2019).

Consumer – a consumer is a person who uses city services (Castelli et al., 2013; Şanta, 2022).

Indicator – an indicator is used to measure a city's performance and it can encompass a variety of dimensions (OECD, 2020; NIST, 2022).

Location – the location where a service is provided to citizens. A location can be an airport, roads, hospitals, etc. (Gil-Garcia et al., 2015).

Stakeholder – A stakeholder is described as a person, group, or organisation, with an interest in or ability to contribute to smart city development. Citizens, energy suppliers, non-profit organisations, and others are examples of stakeholders (Bankauskaite, 2019).

2.5.2 Information Concepts

This section selects concepts from existing EAs for smart cities that describe applications and data.

Application back-end – applications that provide support to deliver services to citizens. Application that process information and end users do not interact with them (Kuk and Janssen, 2011).

Application front-end – the front end application includes interfaces used by end users (citizens or business) to have access to cities services (Kuk and Janssen, 2011).

Data – data is collected, processed, structured, and analysed, to be useful and provide services to end users (Jin et al., 2014).

Document – document or record that are produced by public agencies and private sectors and their handling and access to end users are based on their sensitivity level (Anthopoulos and Tsoukalas, 2006).

Domain application service – domain application service includes services provided in a specific domain such as health, energy, transport (Gaur et al., 2015).

Middleware – middleware is a software that assists in the communication between applications and operating systems (Kuk and Janssen, 2011).

Monitor application – is an application that monitors specific events through a variety of devices around a city (e.g., noise, air quality) (Hefnawy et al., 2016).

Notification - notification is a message that is created by devices (e.g., sensor) based on city events (e.g., air pollution, temperature) (Gaur et al., 2015).

Public and private data – data collected in a smart city can be classified into public and private data, and they must be processed, taking into account the requirements of each category of their sensitivity (Anthopoulos and Tsoukalas, 2006).

Software service – used to develop and maintain applications and services to citizens (Santana et al., 2017).

2.6 Metamodels for Smart Cities

A model is an abstraction of real-world phenomena and a metamodel defines a domain-specific modelling language to describe a model, whereas a modelling language is defined by abstract syntax, concrete syntax, and semantics (Karagiannis et al., 2019). The metamodel plays an important role, as it reduces complexity when creating conceptual models. It defines the concepts, their relationships, rules, and constraints (Bork, 2018).

EAs can be large models mainly due to concepts necessary to represent, relationships and their combinations that need to be captured. EA metamodels can assist enterprise architects to handle complexity in a smart city context and to be used as a common language to facilitate communication with other stakeholders (Chiprianov et al., 2014).

Rosique et al. (2017) created a metamodel to aid in the modelling of a smart city system utilising domain-specific concepts. The metamodel's goal is to offer a simple method for modelling any system in a smart city, regardless of platform, communication protocols, etc.

Abu Matar and Mizouni (2018) have developed a reference architecture metamodel to be used as a starting point for smart city projects. The metamodel provides variability mechanisms to allow instantiations of different smart city software architectures. SmartCityRA is a modular framework and can be extended. The framework contains nine views capability, participant, service, data, business process, application, analytics, place, and infrastructure. The metamodel categorizes variation points as diverse stakeholders, diverse sensor types, diverse data sources, platform variability, geographical distribution, different services and business processes, provisioning variability, and deployment variability.

Badii et al. (2011) have developed a comprehensive and open ontology for smart city named Km4City. Examples of domains covered by this ontology are weather, cultural heritage, public structures, mobility, events, etc. The model shows the classes and relationship between them. In order to facilitate the use of private and open data, Km4City provides a unique point of access for interoperable data of a city through web and mobile applications.

Bastidas et al. (2022) propose an extension of an EA metamodel to address the lack of domain specific concepts in smart cities. The study focus on concepts to provide support of city services managements for instance domains, city services, quality of life and their information systems (i.e. API, dashboard, quality of application services) in alignment with city goals and objectives.

Enterprise Modelling Languages

Some languages like UML and Business Process Model and Notation (BPMN) are classified as general-purpose languages (Bork, 2018). In this current research, UML was used to model the metamodel and will be presented in Chapter 4. However, to represent the metamodel, it is necessary to use an enterprise modelling language. There are currently a variety of EM languages which are graphical or textual languages used for visualizing, specifying, constructing and documenting artefacts (Chiprianov et al., 2014). Barat et al. (2016) provide a study that reviews enterprise modelling techniques, focusing on support decision-making and identifies specification and analysis requirements such as aspects (why, what, how and who), socio-technical characteristics, visualization and analysis. Bork (2019) conducted a survey of modelling language specification techniques aiming to analyse published standard modelling language specifications. Eleven specifications were found and analysed.

2.7 Summary

In this chapter, we identified gaps in the literature, in relation to data lifecycle modelling in a smart city context. This study first identified the difficulties and needs associated with data processing in the context of smart cities.

Following that, we identified and analysed past studies in order to identify challenges related to data lifecycle modelling. This review assisted us in identifying the elements required to model this tool. Furthermore, an analysis of EA frameworks was performed, identifying concepts used to model a data lifecycle in the business and information layers.

During the review, it was discovered that different data lifecycles are required to suit the various data processing requirements. In addition, it was observed that the majority of architectures provide a high level of abstraction in their modelling. As such, they model building blocks, but they lack concepts representation and linkages between them. Moreover, domains are used to represent city services. A framework for a smart city must be flexible in order to accommodate domain-specific requirements.

Following the completion of the literature review, problems with this tool's modelling were identified and addressed by the research questions of this study.

We demonstrated that the elements required to model data lifecycles are not included in the EAs studied during the literature review. Given that smart cities have several data processing challenges, it is evident that their inclusion in architectures is necessary to give stakeholders a comprehensive understanding of data processing so they can provide services to citizens and manage data more effectively as demonstrated in this chapter.

We can infer from the results of the literature review that these models' need to be improved. As a result, the main purpose of this current study is to create a metamodel that will provide concepts to model the variations in data lifecycles required to handle various types of data.

Chapter 3

3. Research Methodology

Research methodology offers a guide on how to conduct research, therefore following a methodology is fundamental to guarantee the quality of results (Johnston, 2014).

The choice of a methodology is an important part of research, as it will be a guide on how to conduct research, therefore, it is necessary to take into account the objectives and requirements of a research (Basias and Pollalis, 2018).

In this chapter, the importance of using a research methodology in conducting research will be addressed. We will discuss different types of research methodology available, qualitative and quantitative, and their main characteristics. Behavioural and design science are also discussed, as they are paradigms used in the Information Systems field.

Subsequently, the chapter presents design science as the methodology chosen to serve as a guideline to conduct this current study, describing the reasons for this choice. The final part of this chapter describes the steps and different methods followed by this study in order to answer the research questions formulated in Chapter 1.

3.1 Importance of Research Methodology

Research is the act of investigating a subject to gain knowledge in order to solve an identified problem. On a daily basis, there is a need to find answers to several problems and a necessity to understand the world around us (Creswell, 2015; Hancock and Algozzine, 2017).

However, conducting research may not always be an easy task, as research is, not only the act of answering questions, but finding answers in a reliable way. Therefore, to conduct formal research, it is necessary to use systematic actions (Kothari, 2004). The best way to conduct research is using a methodology, which will give a researcher direction on how to best conduct their research, thus adding credibility to their work (Basias and Pollalis, 2018; Hancock and Algozzine, 2017). Research consists of several steps (Creswell, 2015; Hancock and Algozzine, 2017):

- what will be studied
- how it will be studied
- how data collection will be done
- how data analysis will be done
- how to evaluate the results
- dissemination of findings

There are several ways to conduct research and the use of a methodology provides a researcher with a guideline so they do not deviate from the focus of their work and the same time have a plan on how to answer research questions to solve a specific problem and to produce valid and reliable findings (Creswell, 2015).

3.2 Research Methodology Requirements

After identifying the research problem followed by development of research questions (chapter 1), the next step was to select a research methodology that would meet the research objectives of this current study.

To conduct this research, it is necessary to select a methodology that supports the creation of an IS artifact, which in the case of this research, is a metamodel. As the objective of this research is to solve a problem faced by practitioners, the chosen methodology therefore needs to provide techniques to gather information from literature and practice.

An appropriate methodology for this research should also provide methods for empirical and rigorous evaluation of an artifact. Taking these factors into consideration, Figure 10 illustrates the research methodology requirement for this current study.

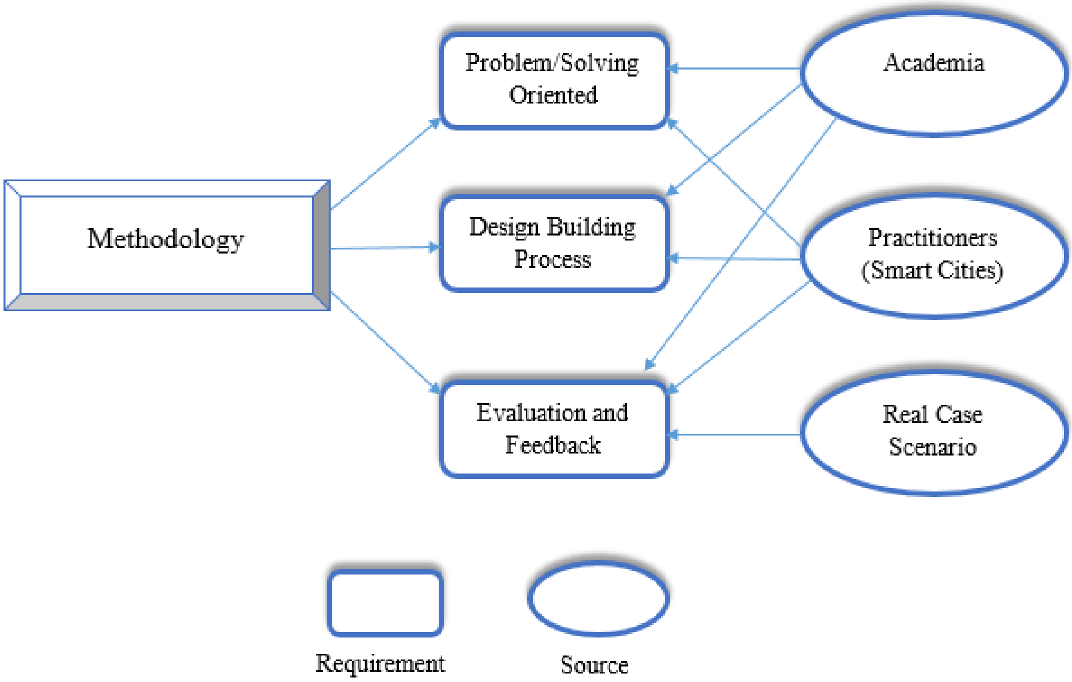

Figure 10 Methodology Requirement

3.3 Methodologies for Research Guidance

As mentioned previously, when conducting research, it is necessary to follow a methodology to guide the steps of research and to ensure the quality of findings (Kothari, 2004). There are several types of existing methodologies and they have different purposes (Hancock and Algozzine, 2017).

In order to carry out research, it is essential to understand the existing methodologies in order to choose a methodology that is appropriate to guide the stages of research.

Research methodologies are classified into qualitative, quantitative, and mixed research. This section describes methodologies, and also behavioural and design science, as they are identified as important paradigms in the Information Systems field.

3.3.1 Qualitative Research

This methodology tries to understand a phenomenon by taking into account people's experiences and their environment (Creswell, 2015). It uses an exploratory approach and its research questions focus on what or how certain phenomena occurs (Hancock and Algozzine, 2017). Data collection is done through interviews, documents, focus groups, and observations focusing on non-numerical data. This methodology uses a small sample of individuals to collect their data to be analysed (Basias and Pollalis, 2018).

Qualitative research has many approaches, and the most common are ethnography, action research, phenomenological studies, grounded theory, and case study, which are detailed below.

Ethnography – is a type of qualitative research that investigates cultural groups. A researcher immerses themselves in a group to describe its lifeways. The data collection is primarily based on observations conducted by the researcher. Interviews can also be conducted with people who have great knowledge about a group or culture. The main objective of this approach is to have a holistic view of a specific group's behaviour, language, and interactions (Hancock and Algozzine, 2017).

Action research – is an approach where a researcher and a client work together to find a solution to a practical problem. This approach is conducted through a cycle, where a plan for a change is made, action is taken, and observation is made to assess the consequences of specific actions. A reflection is conducted to analyse the process carried out and then a new cycle is started. In other words, learning of the phenomenon is gained through action (Koshy et al., 2011).

Phenomenological studies – this approach studies people's experiences regarding a phenomenon. The main goal is to gather information regarding people's lived experiences to help others to learn from these experiences. Data collection is mainly obtained through interviews (Hancock and Algozzine, 2017).

Grounded theory – it is a systematic qualitative research approach that creates theories that are based on gathering and analysis of data (Hancock and Algozzine, 2017). This approach focuses on the development of theories and not testing them (Chun Tie et al., 2019). Multiple sources are used such as interviews, focus groups, documents, etc.

Case Study – it is an approach used to study a phenomenon in its real-world context. This approach is used when the phenomenon studied is contemporary and to provide an in-depth understanding of it. Multiple sources can also be used in this approach such as documents, interviews, and observations. It can include single or multiple case studies (Yin, 2014).

3.3.2 Quantitative Research

Quantitative research methodology is a systematic and empirical investigation of a phenomenon that focuses on numerical data (Hancock and Algozzine, 2017). Usually, this methodology analyses a large amount of quantitative data, where it tries to investigate relationships between variables, test hypotheses, and analyse the existence of cause-and-effect relationships between variables (Creswell, 2015). Unlike the qualitative methodology, quantitative methodology uses surveys, and questionnaires as a form of data collection in the format of closed questions (Basias and Pollalis, 2018).

Another difference between the methodologies is the size of the sample; in quantitative, the sample size is larger than in qualitative. Data analysis occurs using statistics tools. Quantitative research is objective, and a researcher's opinions or feelings do not influence the results (Hancock and Algozzine, 2017). The main quantitative methods are presented below:

Experimental Research – in this quantitative method, a researcher plans experiments to test ideas in order to find out if it has any influence on the final result. This type of method is used when a researcher wants to find out if there is a cause and effect between independent and dependent variables.

Experiments are conducted with controlled conditions controlling an environment in order to investigate the relationship between variables to test or validate a hypothesis (Creswell, 2015).

Non-experimental Research – this quantitative method is used when there is no manipulation of an independent variable or other conditions in a research. Researchers use this method to identify and measure the degree of relationship between variables, identifying how much one variable influences another and the results are primarily descriptive (Creswell, 2015).

3.3.3 Mixed Research

Mixed research is a method, where both quantitative and qualitative methods are used to answer questions related to one research problem. It is used when quantitative and qualitative data is available (Creswell, 2015). Using both methods can provide a better picture of a research problem and provide more in-depth results. However, with the adoption of different methodologies, a researcher should merge, connect, build, and embed data in order to conduct a research, therefore, requiring specific skills from researchers (Hancock and Algozzine, 2017).

Mixed research allows researchers to use a variety of methods, promoting exploration of different perspectives and enabling a better understanding of a problem and its findings (Kothari, 2004).

3.3.4 Behavioural Science and Design Science Research

Two distinct, but, complementary, paradigms have been used to conduct research in IS: behavioural science and design science (Hevner et al. 2004). Behavioural science tries to develop theories to explain and describe human behaviour or organizational behaviour encompassing analysis, design, implementation, and management, and it is used in information systems. It originated in natural science research starting with a hypothesis that a researcher proves right or wrong (Hevner, and Chatterjee, 2010).

On the other hand, design science provides the necessary steps to conduct research in order to develop an artifact to solve a problem in a real application environment (Hevner et al. 2004; Peffers et al. 2007). One of the goals of design science research is to solve "wicked" problems in a challenging problem domain inspired, motivated, or informed by practice. "Wicked" problems are described as unique and can be considered a symptom of another problem; they have a lack of definitive formulation, unstable requirements, and constraints, and complex interactions among subcomponents of problem and solution (Rittel and Webber, 1973; Hevner and Chatterjee, 2010). The characteristics of "wicked" problems can be applied to the challenges faced by data management in smart cities (EMC, 2015; Chen et al. 2014; Chauhan et al. 2016).

3.4 Research Methodology Selection for Guiding this Research

Section 3.2 presented different research methodologies and their characteristics. Design science research methodology has been chosen as a guideline to conduct this current research and the reasons are explained as follows.

First, design science is a well-established research methodology in the IS field, where it creates innovative artifacts to solve human problems, hence contributing new knowledge to the knowledge base in the IS field (Hevner et al. 2004). Secondly, the objective of this research is to model variations of data life cycles in EA in the smart city domain.

Thirdly, design science defines the steps of a process to design, evaluate and demonstrate an artifact, offering flexibility in the use of methods and tools while guaranteeing the quality of an artifact. Table 5 outlines the arguments used to support the selection of the design science research methodology used to conduct this research.

Table 5 Summary of research methodologies

Research Approach	**Steps**	**Problem Solving**	**Design Process Building**	**Research Outcome**
Action Research	- Plan - Act - Observe - Reflect	Yes	No	Solution for a specific problem
Grounded Theory	- Data Collection - Analysis - Theoretical sampling - Theory building	No	No	Theories
Design Science	- Problem & motivation - Define objective - Meta design - Artificial evaluation - Design	Yes	Yes	Constructs Models Methods Instantiati ons

	practice - Demonstration - Naturalistic evaluation - Communication			
Case Study	- Plan - Design - Prepare - Collect - Analyse - Share	No	No	Phenomenon investigation

3.5 Design Science Research Methodology

DSRM is commonly used in the Information Systems field. Hevner et al. (2004) demonstrate that, in order to solve a problem in the Information Systems environment, a business needs to develop theories and artifacts using a knowledge base, where it has foundations (theories, models, frameworks) and methodologies (data analysis, techniques, measures), which it is acquired from the development of theories, artifacts and accessing using field study, simulation, case study, just to name a few. One of the goals of design science research is to solve a problem that is inspired, motivated, or informed by practice.

Peffers et al. (2007) defined the steps to implement the design science research, as identify a problem and motivate, define objectives of a solution, design, and development, demonstration, evaluation, and communication, with the advantage of iterations.

This current research uses the methodology proposed by Ostrowski and Helfert (2012), which improved the original process of its gathering and modelling information in collaboration with practitioners. The steps of the utilised methodology can be seen below in Figure 11. Details of each step of the selected framework applied in this research will be presented in the next section.

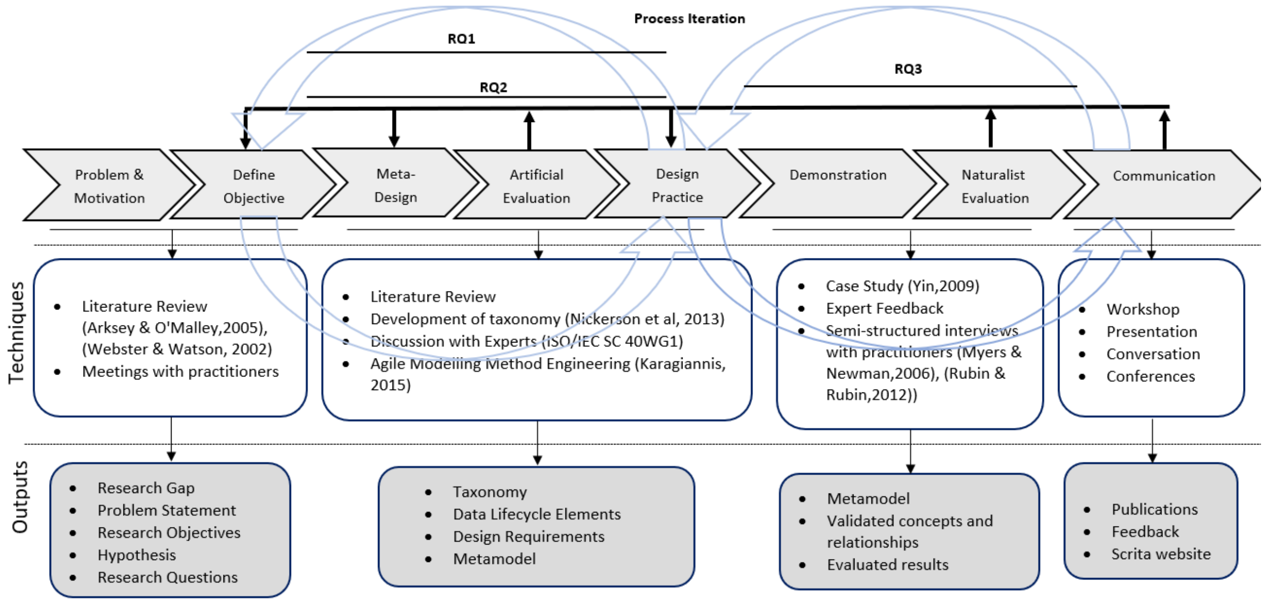

Figure 11 Design Science Methodology followed by this research (Ostrowski and Helfert, 2012)

3.6 Design Science Application in this Thesis

Considering the need to investigate and understand why data influences the choice of activities and phases in a data lifecycle, this research is going to use qualitative methods. Moreover, the current researcher has chosen to follow the design science research methodology (DSRM), which is a problem-solving paradigm that offers flexibility by allowing the use of different methods to conduct research. This section describes the steps of the methodology presenting methods and techniques used to address the research questions of this research.

Problem & Motivation and Define Objective

In the first two steps – Problem & Motivation and Define Objective, a scoping review was conducted in order to obtain knowledge in the data management area and to find a research gap. The researcher followed Arksey and O'Malley's (2005) framework to obtain an overview of the data management field in the smart city context, which assisted in finding a problem statement.

Subsequently a systematic literature review was conducted, with the purpose of identifying issues in data lifecycle modelling and therefore refine the problem statement and development of hypothesis and research questions. The systematic literature review followed the guidance and concept matrix provided by Webster and Watson (2002). The research gap is presented in Section 1.2 of this thesis. Problem identification and motivation are identified in Section 1.3 with research objectives presented in Section 1.5.

Meta design, Artificial Evaluation & Design Practice

This part of the research covers the design phase of an artifact to answer research questions 1 and 2. The purpose of this research is to model data lifecycle variations within EA in the smart city domain, therefore it is necessary to identify elements of a data lifecycle, hence a systematic literature review was conducted to identify them. It also conducted an analysis of sixty-three data lifecycles from different domains, of which eight were developed for smart cities. The reason behind this is the existence of several data lifecycle models, therefore it was indispensable to analyse them and understand why and when these variations are required. The details of systematic literature review can be found in Section 2.1.

In addition, a data taxonomy was developed to provide an understanding of different types of data and their requirements, as phases and activities of a data lifecycle change based on the type of data, which needs to be processed. The data taxonomy was developed, taking into account the specifications of the smart city domain.

To develop the taxonomy, I followed a methodology proposed by Nickerson et al. (2013), which proposes development by iterations, and after each iteration, it is checked if the taxonomy has achieved ending conditions. To identify new objects, I used coding and categorization of data using the inductive approach, proposed by Thomas (2006), which consists of identifying keywords and labelling them into categories.

Meetings with practitioners from Cork City Council, the Central Statistics Office, and Belfast City Council were conducted, not only with the purpose of validating findings in the literature, but also to obtain insights and challenges faced by them; as the Ostrowski and Helfert (2012) model emphasizes the importance of working with practitioners in conjunction with a literature review. Meetings with experts from the ISO working group were conducted in order to obtain their feedback and point of view, regarding the data classification phase.

Agile Modelling Method Engineering

In order to improve data lifecycle modelling in enterprise architectures in the smart city domain, the solution proposed was to develop a metamodel. A metamodel is used to define constructs in enterprise architecture, and describe relationship between these constructs and attributes to describe them. As design science research offers flexibility to choose techniques to develop an artifact, I used the agile modelling method engineering proposed by Karagiannis (2015; Karagiannis et al., 2019), to develop the metamodel, which is a paradigm to address changes in modelling requirements. The phases of AMME cycle are depicted in Figure 12 and described below.

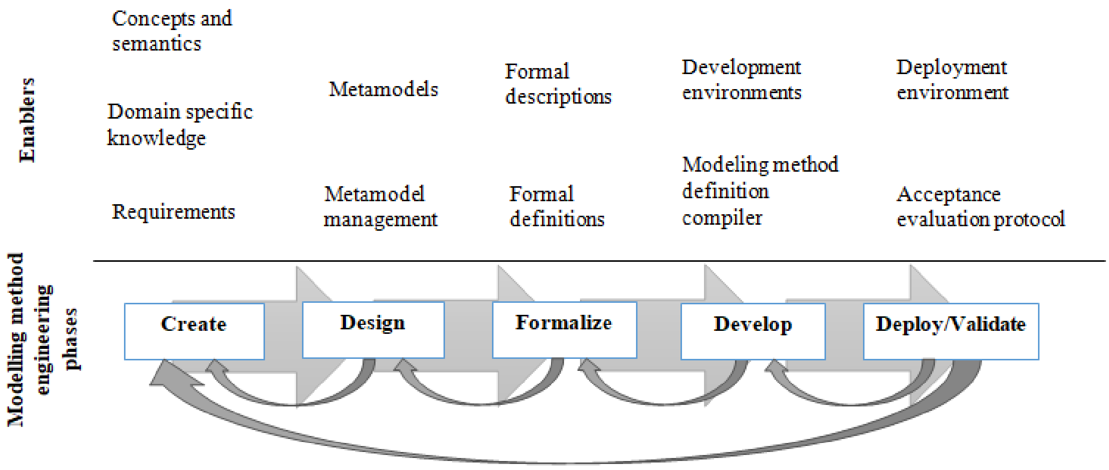

Figure 12 Agile Modelling Method Engineering (Karagiannis, 2015)

The method is composed of six iterative phases to create design, formalise, develop and validate a metamodel. This modeling method ensures the development of models taking into account certain aspects such as gradual changes in requirements for modeling in a given domain as enablers are presented at the top and they will be used during the development phases.

Create phase focuses on knowledge acquisition and requirements elicitation. Design phase focuses on the metamodel which will be developed, paying attention to language grammar, notation, and functionalities. Formalise phase formalizes the outcome from previous phase. Develop phase maps the concepts identified to ArchiMate language. The last phase focuses on feedback from stakeholders and evaluation of the metamodel developed.

Model Driven Engineering

Model-driven engineering (MDE) emerged as an approach to minimize problems encountered in software development where several projects encountered issues for instance, software not meeting customer specifications, and software developed with low quality, among others (García Díaz et al., 2014).

This paradigm suggests the use of software models in order to separate the functionality of a system that is being developed and its implementation, thus increasing the quality of software development. The objective of using these models is to increase level of abstraction and improve the collection of requirements and specifications on customer side, reducing errors in this phase (Bezivin, 2004).

Silva (2015) states that a model is a representation of a real-world element in a specific domain and can be used to improve communication between stakeholders and to create a model, it is necessary to create a metamodel, which contains elements of a model, its relationships, and rules.

This thesis follows the Model Driven Engineering approach to develop the proposed metamodel. Figure 13 illustrates the 3+1 architecture defined by the OMG (Bézivin, 2004).

The M0 layer represents the real system, which is modelled on the M1 layer. This model is in conformance with its metamodel defined on M2 layer. Finally, the metamodel conforms to the meta-metamodel on layer M3, which conforms to itself.

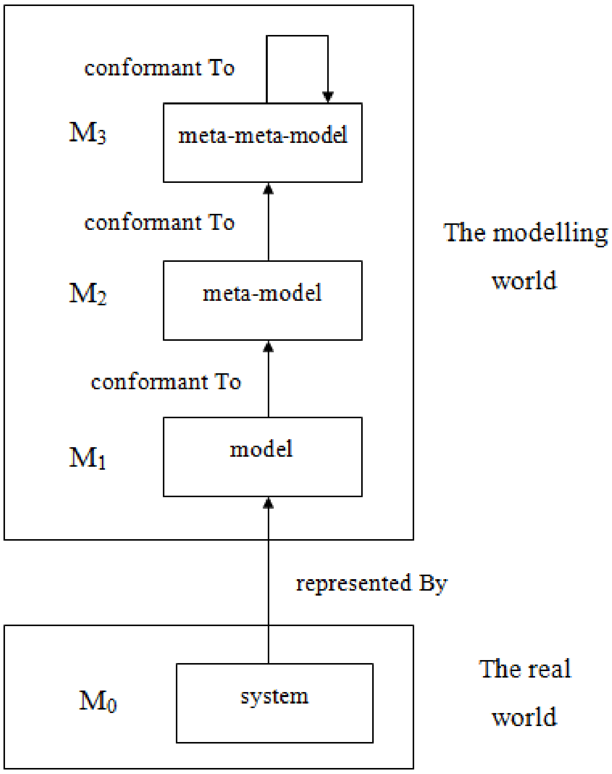

Figure 13 3+1 Model Driving Architecture (Bézivin,2004)

Data Taxonomy

To understand how data processing requirements, influence phases and activities of a data lifecycle, this study developed a data taxonomy in order to understand data specifications from the smart city domain. The development followed the approach suggested by Nickerson et al. (2013) and literature was used to identify new objects following an approach proposed by Thomas (2006) that proposes usage of coding and categorization. Figure 14 depicts the taxonomy development method proposed by Nickerson et al. (2013).

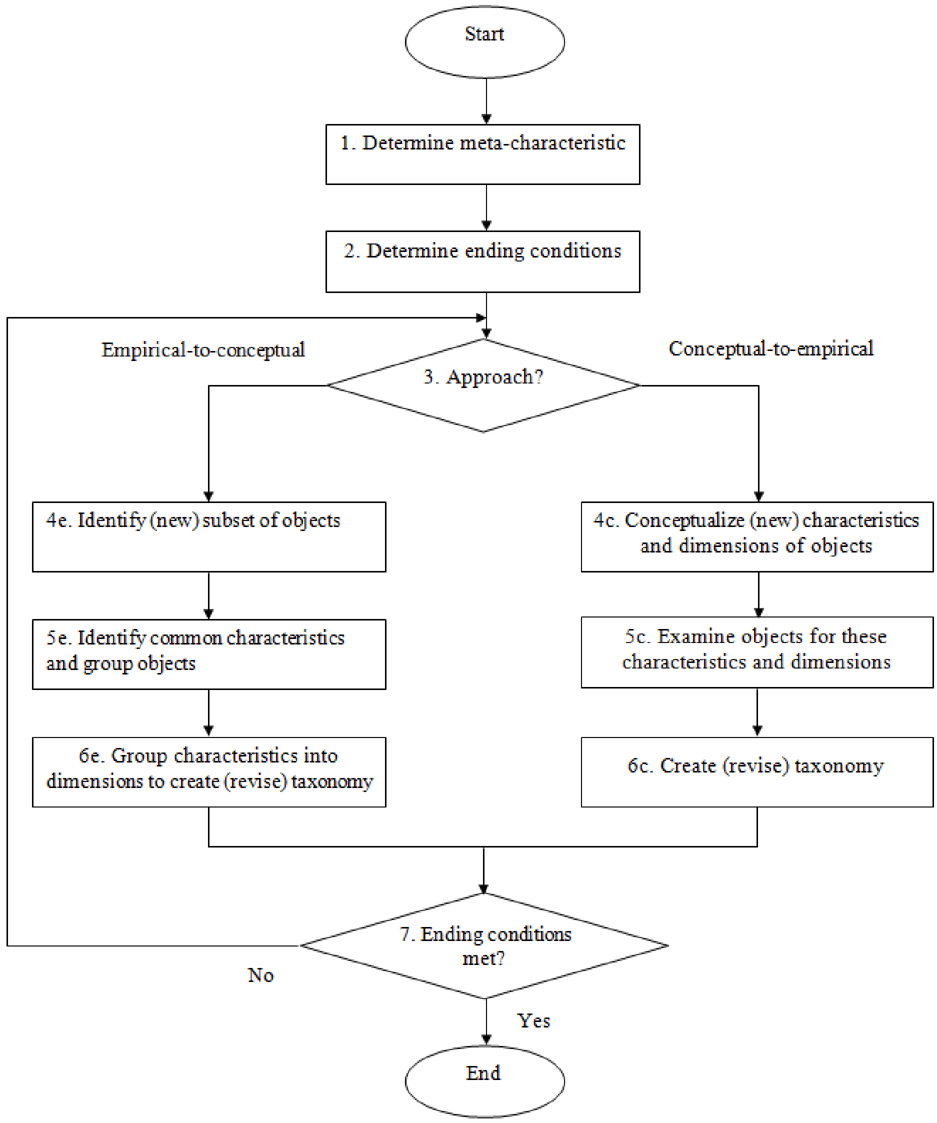

Figure 14 Taxonomy development (Nickerson, 2013)

The literature review identified factors that must be taken into account when data is processed which may vary according to requirements that must be met to process data, causing phases and activities of data lifecycles to also vary. Several studies identify organizations that define how data will be used by defining their data principles. There are several data principles used by cities and this study uses data principles defined for smart cities (EU2019.FI, no date; Bria et al., 2018; Data and Action, 2019; Office of the Government Chief Information Officer and Department of Public Expenditure and Reform, 2019).

Demonstration and Evaluation

Evaluation is an essential element in the development of an artifact (Hevner, and Chatterjee, 2010). Case studies were conducted to evaluate the artifact in the real world. Chapter 5 presents in more detail the demonstration and evaluation phases of this thesis. This section describes the methods and techniques used in the evaluation phase.

Built-evaluate cycles in this thesis

DSR is composed of two main blocks of activities: built and evaluated (Sonnenberg, 2012). Sonnenberg (2012) also affirms the relevance of evaluating the built phase of an artifact, where there is a need to validate design decisions before developing an artifact. The assessment constructed cycles of this thesis are shown below.

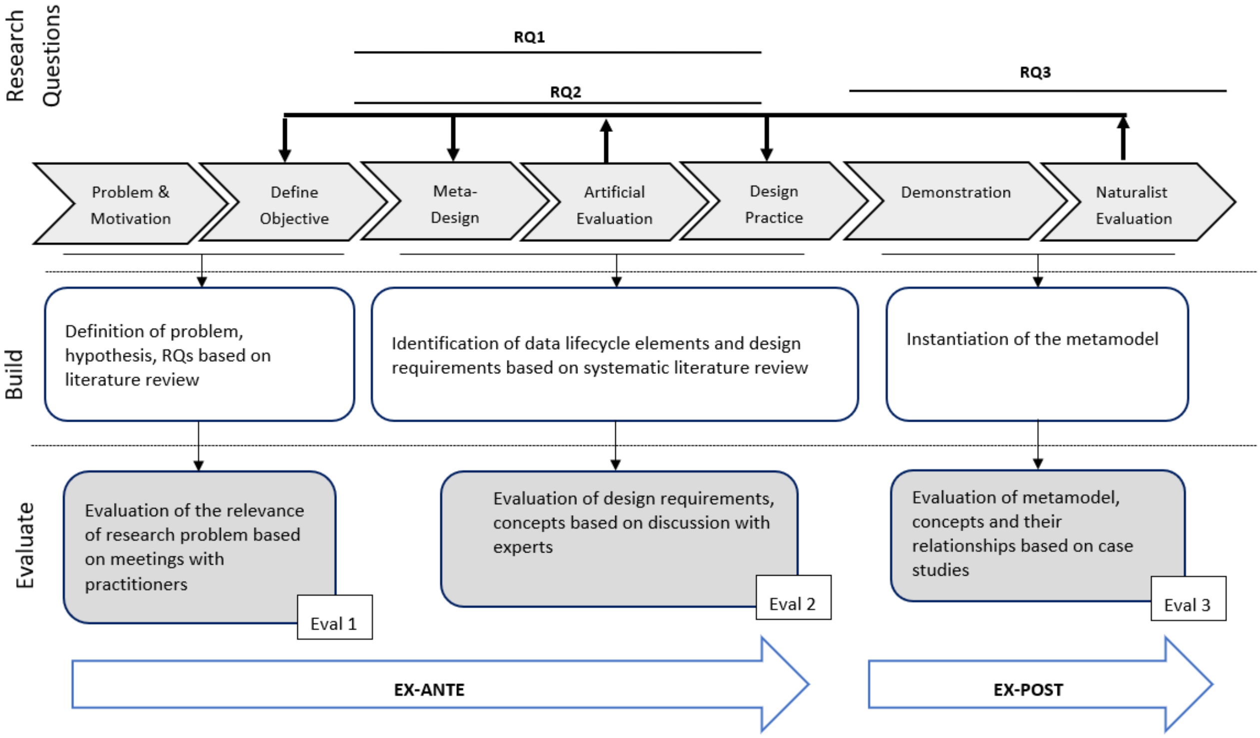

Figure 15 Built-evaluate cycles of this research

Figure 15 shows that the artifact is evaluated in two phases: ex-ante and ex-post. The ex-ante phase evaluates an artifact before it is instantiated. In this study, ex-ante evaluates the problem and design requirements based on a literature review and meetings with practitioners. The ex-post phase evaluates an artifact after it is instantiated. In this study, the metamodel, its concepts, and its relationships are evaluated based on case studies conducted in city councils in Ireland. The steps and methods used in the three evaluations are explained below.

Evaluation 1

It occurs after the identification of a problem, and its purpose is to make an analysis regarding the need to construct an artifact. Therefore, after identifying problems in data lifecycle modelling through scoping review, this was confirmed through a systematic literature review and meetings with practitioners from Cork City Council, Belfast City Council, and the Central Statistics Office.

Evaluation 2

This is an artificial evaluation, in order to evaluate the design of an artifact since it has not yet been instantiated, therefore in this activity the design requirements are evaluated based on discussion with experts focusing on the completeness of the artifact.

Evaluation 3

In the last evaluation, the artifact is used and the aim is to verify its applicability and utility in practice. Therefore, case studies were conducted in Limerick and Cork cities in Ireland. The case studies were conducted to obtain an in-depth understanding of usage of different data lifecycles in the smart city domain and to evaluate the metamodel developed. Data was collected from various sources such as semi-structured interviews, meetings, and city councils' websites.

3.7 Philosophical Stance of the Thesis

With regard to the philosophical component of research, it can be seen from two perspectives: ontology, and epistemology. The first refers to the nature of knowledge, and how we see the world, and the latter refers to how we gain knowledge (Wahyuni 2012; Recker 2005). Taking this into consideration, this research follows an interpretivist approach. From an ontological perspective, knowledge is based on stakeholders' perceptions of reality. On the other hand, epistemological perspective, knowledge is developed through lived experiences. As a result, we cannot state firmly that the findings of this study are the one and only truth, however it focuses on creating an artifact to assist stakeholders.

3.8 Criteria to Assess the Quality of Research Design

Some concepts are used to assess the quality of a research design (Yin, 2014). Therefore, in order to assess the quality of this research design, this current study is applying four tests: construct validity, internal validity, external validity, and reliability as presented below and in Table 6.

- Construct validity is about the correct identification of a set of measures regarding concepts that are going to be studied (Yin, 2014).
- Internal validity is about investigating cause and effect relationships between events (Yin, 2014).
- External validity refers to how results of a study can be generalized independent of research methods used and be applied to other contexts (Yin, 2014)
- Reliability is conducted to ensure that research can be conducted by another researcher, following the same methods, and that the results found will be the same (Yin, 2014).

Table 6 Criteria to assess quality in this thesis

Test	Technique	Used in this research
Construct Validity	- Data triangulation (multiple sources of evidence)	- Case study design and method (Yin, 2014) - Case study research (Hancock and Algozzine, 2017)
Internal Validity	- Systematic literature review - Enterprise architecture guidelines, modelling techniques - Interviews with smart city experts and ISO group	- TOGAF (The Open Group, 2018) - Literature review - Concept centric approach (Webster, and Watson,2002) - ArchiMate (The Open Group, 2019) - Interviews (Myers and Newman, 2007)
External Validity	- Replication in multiple case studies	- Case study (Hancock, and Algozzine, 2017; Yin, 2014)
Reliability	- Modelling method - Case study protocol - Case study database (NVIVO)	- Agile modelling method engineering (Karagiannis, 2015) - Case study (Hancock, and Algozzine, 2017; Yin, 2014) - Qualitative data analysis (Rubin and Rubin, 2005)

Communication

Communication is the last phase of DSR framework. Research must provide a clear contribution, hence it is essential to present the findings to technology-oriented and management-oriented audiences (Hevner et al., 2004). During the course of this study, articles were presented at conferences and in a book. The research was also disseminated at meetings, and poster presentations at events. All these events were valuable opportunities to communicate the study and also to obtain feedback and insights from participants. Moreover, the research group of which I am part has created a webpage to extend the dissemination of our research, therefore I have used this opportunity to share my research findings (Figure 16).


scrita.lero.ie/managing-the-lifecycle/

Enterprise Architecture Management for Smart Cities

A Reference Methodology for Developing and Transforming Public Services

Home Methodology Meta-Modelling Use Cases Research Glossary About Us

The importance of using a Data Life Cycle in the new digital world

The Cambridge Dictionary [1] states that data is "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer".

The digital evolution has bringing a significant transformation in our connected world, as a result of this, a huge amount of data is generated and collected every day from different sources and in order to stay competitive and to help decision making, organizations are always trying to extract value from this vast source of information, therefore, data is being considered the most valuable asset an organization has, consequently it is vital for organizations to know what data they have and classify them based on their value in order to protect them.

A data life cycle (Figure 1) is a framework with all the stages/activities that data has to go through from its creation, processing, analysis, storage and sometimes disposal. Considering the pivotal role which data plays worldwide nowadays, the usage of a data lifecycle is indispensable, as it can assist the organisations to ensure that data is collected and prepared for the intended usage and end users.

Figure 1 – Example of IoT infrastructure with data lifecycle

Source: Jin et al. 2014

Figure 16 Scrita Website

3.9 Summary

In this chapter, we discussed the importance of a research methodology to conduct research. The main characteristics of different types of research methodologies available were presented. The research methodology requirement was outlined to assist in the selection of a methodology to conduct this research. Design Science was chosen as the research methodology to be followed and the methodology framework was discussed. The methods and techniques used to develop the proposed metamodel were identified and detailed. The chapter concludes by discussing the criteria for evaluating the quality of this research design, including validity and reliability. The following chapters describe the thesis's metamodel design, evaluation, and conclusion.

Chapter 4

4. Metamodel Design

The literature review presented in Chapter 2 revealed several drawbacks in data lifecycle modelling. The main problem identified is the tool's high level of abstraction and its inability to represent the reality of data management. Chapter 2 also showed the data management challenges faced by a smart city and how its use of data lifecycles is essential.

This chapter explores concepts about smart cities that are needed to model a data lifecycle in enterprise architecture (EA), which is the objective of Research Question 2. The concepts were identified in the literature and validated during case studies conducted in Ireland. This chapter also shows how design requirements are identified and used to define concepts in the context of smart cities. These concepts are subsequently used to develop a metamodel, which defines how to model variations in data lifecycles in enterprise architectures.

4.1 Overview of Metamodel Development

The research described in this section follows the MOF, which was defined by the Object Management Group (OMG; Bezivin, 2004). The model is divided into four layers (Figure 17). The lowest layer, M0, represents reality; it contains real elements of the world. In this study, M0 represents the domain of smart cities. Above it, layer M1 represents a model of a specific domain. The models in this thesis provide services to citizens. Above M1, the M2 layer represents a data lifecycle metamodel, which describes elements from the previous layer or how end users can model data lifecycles in EAs for the smart city domain. Finally, the top layer, M3, contains MOF, which is a standard proposed by the OMG for creating metamodels that are instances of layer M2 (OMG, 2016).

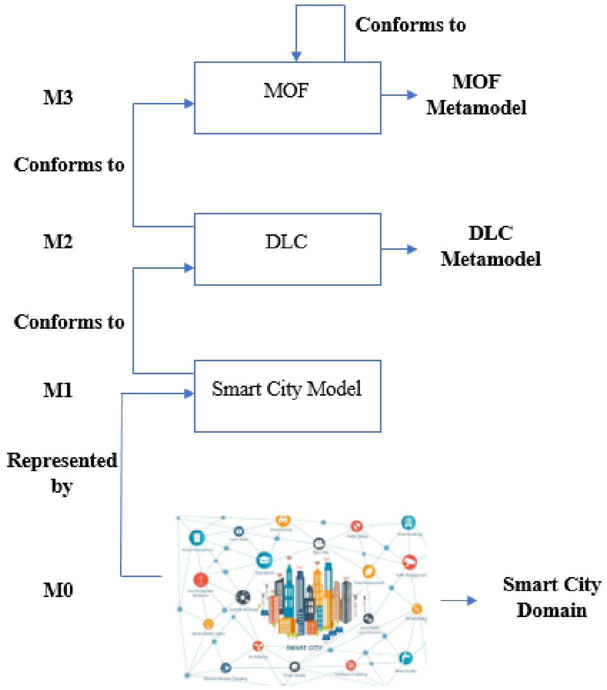

Figure 17 Proposed metamodel based on 3+1 architecture by OMG (OMG, 2016)

Figure 18 depicts the tasks and phases that this study uses to define concepts and build a metamodel. This method of building a metamodel, which was proposed by Karagiannis (2015), is called Agile Modelling Method Engineering (AMME).

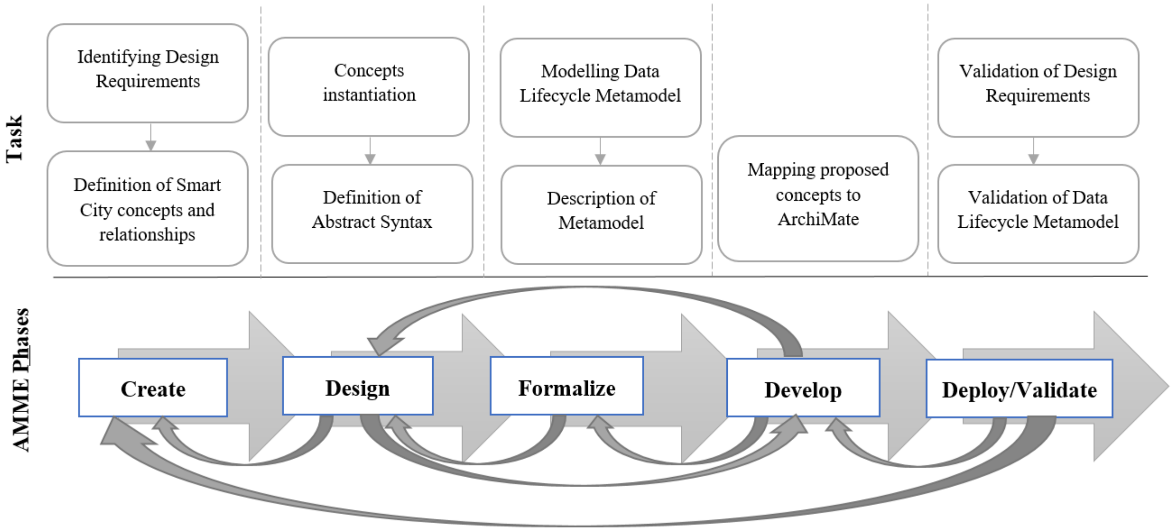

Figure 18 Metamodel Development Method (Karigiannis, 2015)

4.2 Phase I: Create

The create phase identifies concepts utilising a taxonomy created from literature, as well as design requirements extracted from literature to capture specific knowledge of the smart city domain. The sections that follow present the taxonomy and the design requirements that were used in this thesis.

4.2.1 Data Taxonomy

As mentioned in Section 3.6, a data taxonomy was developed using an approach suggested by Nickerson et al. (2013), and literature was used to identify new objects following an approach proposed by Thomas (2006) that involves coding and categorization. The taxonomy was used to identify and categorize the concepts necessary to model data lifecycles in the proposed metamodel. The concepts processing type, category, sensitivity, data type and processing location were identified using the taxonomy. The taxonomy, represented in Figure 19, required five iterations to develop. The numbered steps below illustrate how the method in Figure 14 is applied.

Iteration 1:

Step 1: Meta-characteristic: stakeholders involved in the data management process and access type.

Step 2: Ending condition: The method ends when objective and subjective conditions have been met.

Objective condition: No new dimension is added in the last iteration and subjective conditions from table provided from Nickerson et al. (2013), where contains a list of questions to guarantee that dimensions are concise, robust, comprehensive, extendible and explanatory.

Iteration 2:

Step 3: Approach: empirical to conceptual.

Step 4e: We identified the stakeholders from the literature.

Step 5e: We identified the access type the stakeholders can have to manage data.

Step 7: Ending condition: Only one dimension was created, therefore it is necessary to have more iterations.

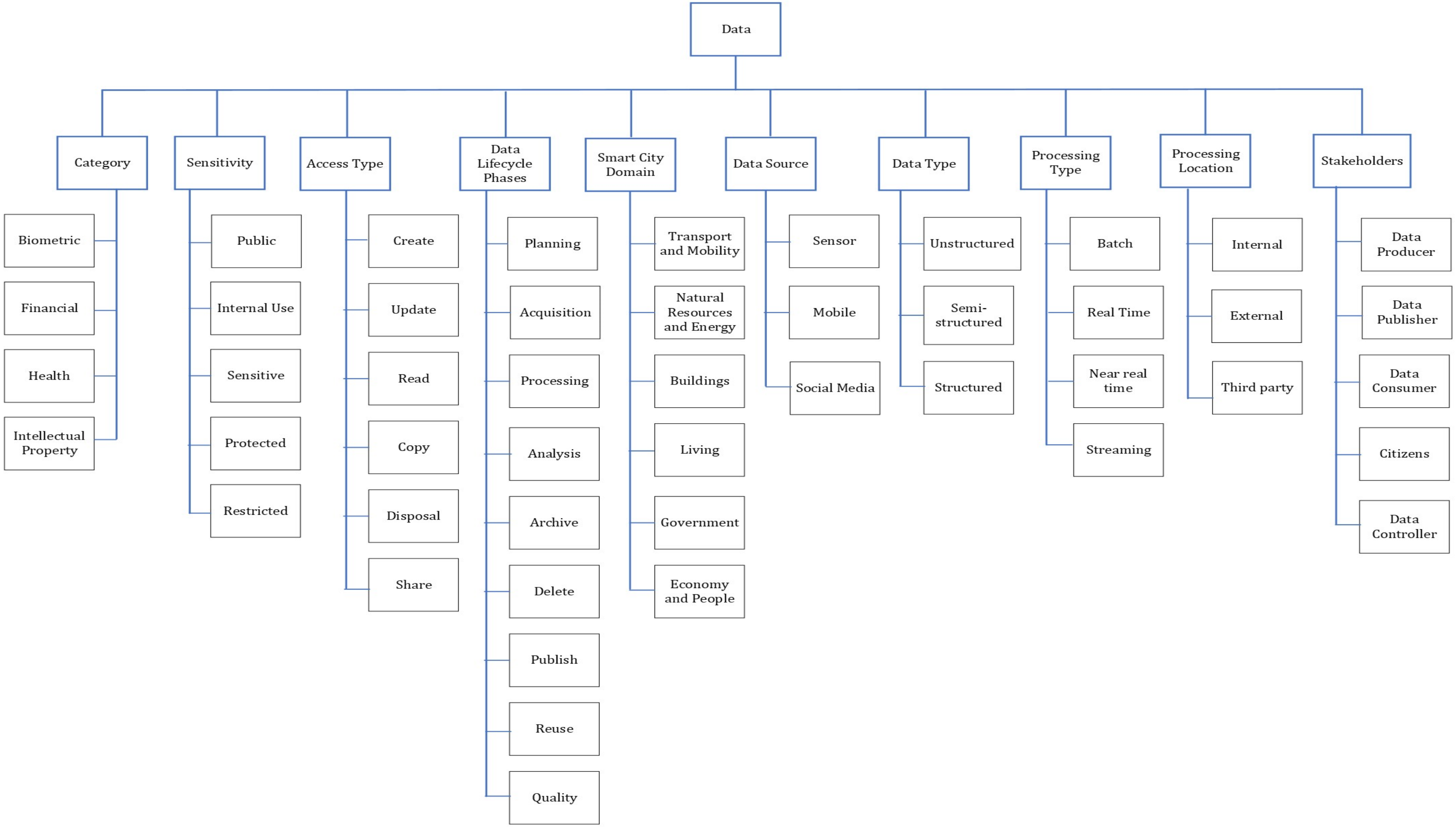

Figure 19 Data Taxonomy

4.2.2 The need for design requirements

The first step in modelling a data lifecycle in enterprise architecture is to identify its requirements. This study identified requirements in the literature using the approach proposed by Chantra et al. (2015), which focuses on what an artifact must allow at the end of its development; in this view, requirements are used to define functionalities that an artifact must have. In the case of this study, requirements define functionalities that the metamodel must meet. Table 7 outlines the identified requirements.

Table 7 Design Requirements

Design Requirement	**Concept**	**Supporting Source**
1. Provide exclusive concepts to manage a data lifecycle	- Objective - Phases/Activities Planning Collection Classification Anonymization Usage Storage Publication Destruction	Shah et al. (2021), Christopherson et al., 2020) Lim, et al. (2018), Paskaleva, et al. (2017), Attard,et al. (2016), Liu, et al. (2017), Alshammari and Simpson (2018), Gharaibeh et al (2017), Faundeen and Hutchison (2017), Lněnička and Komárková (2019), Ireland. Department of Public Expenditure and Reform. (2018),

		Alshammari and Simpson (2018b), Charalabidis (2018), Weber and Kranzlmüller, (2019), Erl et al. (2016), Elmekki et al. (2019), Plale and Kouper (2017), Sutherland and Cook, (2017), Lim et al. (2018).
2. Provide exclusive concepts to manage variability in data lifecycles	- Variation Driver	Rurua et al. (2017), Shah et al. (2021), Christopherson et al., 2020, Abu Matar and Mizouni (2018), Pohl et al. (2005)
3. Provide exclusive concepts to manage data requirements	- Data Type - Processing Type - Processing Location - Data Lifespan	Gharaibeh et al (2017), Ireland. Department of Public Expenditure and Reform. (2018), Shah et al. (2021), Moustaka et al (2018), Christopherson et al., 2020, Erl et al. (2016), British Standard (2017), Ahmad (2022)
4. Provide exclusive concepts to classify data	- Category - Sensitivity - Policies/Regulations	Attard,et al. (2016), Liu, et al. (2017), Alshammari and Simpson (2018), Gharaibeh et al (2017), Lněnička and Komárková

		(2019), Ireland. Department of Public Expenditure and Reform. (2018), Atis (2018), Alshammari and Simpson (2018b), Erl et al. (2016), Elmekki et al. (2019), Plale and Kouper (2017), British Standard (2017), Demchenko et al. (2018), Eurocities. (2019), Sutherland and Cook, (2017), Lim, et al. (2018), Ahmad (2022)

Design Requirement 1: Concepts to manage a data lifecycle

Data is the main resource that a city can use to become smart and offer better services to its citizens. However, there is a deluge of data nowadays, and cities are the main producers of data. Processing data is a complex task in a smart city due to several factors. As explained in Chapter 2, data lifecycles can help stakeholders reduce this complexity by dividing data processing into phases and activities, thereby achieving better planning, understanding, and requirements needed for processing data (Arass et al., 2017; Sinaeepourfard et al., 2016; Weber and Kranzlmüller, 2019).

Concepts identified for managing a data lifecycle include objectives, phases, and activities. This data management tool has an objective – the purpose for which it is carried out – and it has to provide a specific outcome in the end (Alshammari &

Simpson, 2018; Christopherson et al., 2020). A data lifecycle is divided into phases, which in turn are divided into activities (Shah et al., 2021; van Veenstra & van den Broek, 2015). Phases of this tool vary with the different requirements for processing data as activities of each phase (Alshammari & Simpson, 2018; Shah et al., 2021).

Design Requirement 2: Concepts to manage variability in a data lifecycle

As discussed in Chapter 2, data lifecycles vary due to several factors; therefore, it is necessary to have a concept that identifies this variability. Because a smart city offers different services to its citizens, it must have different data lifecycles to process data. The stages and activities of the data lifecycle vary with the requirements that data processing must meet. The variance driver is a foundational concept for identifying this variability. It allows a city to identify requirements that influence variations needed in data lifecycles to process data, produce a final outcome and to meet city data strategy.

Design Requirement 3: Concepts to manage data requirements

A city offers a variety of services in different domains, such as mobility and transportation, the environment, and healthcare. A smart city collects data about these different types of services, but it must process them differently and at different locations. Data type, processing type, and processing location are concepts required to model data processing in a smart city.

Design Requirement 4: Concepts to classify data

In data processing, classifying data into categories is essential in order to use and protect data more efficiently. Because cities produce a huge amount of data, it is necessary to identify concepts that stakeholders can use to classify the data. It is necessary to take into account sensitivity of data as well as regulations that must be met during data processing. Due to these requirements, this study added the concepts of category, sensitivity, data lifespan, and regulations to data lifecycle modelling in the smart city domain.

4.2.3 Definition of Concepts

This section identifies concepts and their relationships to formalise the design requirements that were identified in Section 4.2.1. These concepts and their relationships are essential for modelling variations of data lifecycles. The design requirements are used to derive concepts that in turn can be used as features of an artifact (Meth et al., 2015). The concepts are described in Tables 8 through 18, and each table describes the concepts related to specific design requirements, their relationships, stakeholders interested in the concept, the justification behind the choice of concepts, and supporting sources from literature.

Table 8 Concept - Objective

Design Requirement	DR 1. Provide exclusive concepts to manage a data lifecycle
Concept	Objective
Relationship(s)	▪ meets (objective, data lifecycle): The relationship between data lifecycle phases that meets an objective.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	Data is processed to achieve a purpose therefore a stakeholder should be able to identify if a data lifecycle is meeting the target objective of processing data.
Supporting Sources	Shah et al. (2021), Christopherson et al., 2020) Lim, et al. (2018), Paskaleva, et al. (2017), Attard,et al. (2016), Liu, et al. (2017), Alshammari and Simpson (2018)

Table 9 Concept – Phases

Design Requirement	DR 1. Provide exclusive concepts to manage a data lifecycle
Concept	Phases
Relationship(s)	▪ processed through (data, phases): A relationship between data that is processed through phases.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should define phases of a data lifecycle necessary to process data for a specific outcome.
Supporting Sources	Ireland. Department of Public Expenditure and Reform. (2018), Alshammari and Simpson (2018b), Charalabidis (2018), Weber and Kranzlmüller, (2019), Shah et al. (2021), Christopherson et al., 2020)

Table 10 Concept – Variation Driver

Design Requirement	DR 2. Provide exclusive concepts to manage variability in data lifecycles
Concept	Variation Driver
Relationship(s)	▪ impacts (variation driver, phases): A relationship between a variation driver that impacts on phases and activities of a data lifecycle
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should be able to identify the variation drivers that impact on data lifecycle phases and activities.
Supporting Sources	Rurua et al. (2017), Shah et al. (2021), Christopherson et al., 2020, Abu Matar and Mizouni (2018), Pohl et al. (2005), Ireland. Department of Public Expenditure and Reform. (2018)

Table 11 Concept – Data Type

Design Requirement	DR 3. Provide exclusive concepts to manage data requirements
Concept	Data Type
Relationship(s)	▪ belongs to (data, data type): A relationship between data that belongs to a data type. ▪ belongs to (data type, variation driver): The relationship between data type that belongs to a variation driver.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should be able to define which type a certain data belongs to.
Supporting Sources	Moustaka et al (2018), Christopherson et al., 2020, Erl et al. (2016), British Standard (2017), Ahmad (2022)

Table 12 Concept – Processing Type

Design Requirement	DR 3. Provide exclusive concepts to manage data requirements
Concept	Processing Type
Relationship(s)	▪ has (processing type, data): The relationship between data that has a processing type.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should be able to define what processing type data will have.
Supporting Sources	Gharaibeh et al (2017), Ireland. Department of Public Expenditure and Reform. (2018), Shah et al. (2021)

Table 13 Concept – Processing Location

Design Requirement	DR 3. Provide exclusive concepts to manage data requirements
Concept	Processing Location
Relationship(s)	▪ has (processing location, data): The relationship that shows that data has a processing location.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should be able to define where data will be processed in order to provide services to citizens.
Supporting Sources	Shah et al. (2021), Moustaka et al (2018), Christopherson et al., 2020, Erl et al. (2016), British Standard (2017), Ahmad (2022)

Table 14 Concept – Data Lifespan

Design Requirement	DR 3. Provide exclusive concepts to manage data requirements
Concept	Data Lifespan
Relationship(s)	▪ has (lifespan, data): Relationship shows that data has a lifespan. ▪ belongs to (data lifespan, variation driver): The relationship between lifespan that belongs to a variation driver.
Stakeholder(s)	City authorities.
Rationale	A stakeholder should be able to define a lifespan for data.
Supporting Sources	Lim, et al. (2018), Paskaleva, et al. (2017), Attard,et al. (2016), Liu, et al. (2017), Alshammari and Simpson (2018), Gharaibeh et al (2017), Shah et al. (2021)

Table 15 Concept – Category

Design Requirement	DR 4. Provide exclusive concepts to classify data
Concept	Category
Relationship(s)	▪ has (data, category): Relationship shows that data has a category. ▪ belongs to (category, variation driver): A relationship between category that belongs to variation driver. ▪ defines (category, sensitivity): A relationship between category that defines sensitivity.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should be able to define a category which a data belongs to.
Supporting Sources	British Standard (2017), Demchenko et al. (2018), Eurocities. (2019), Sutherland and Cook, (2017), Lim, et al. (2018), Ahmad (2022)

Table 16 Concept – Sensitivity

Design Requirement	DR 4. Provide exclusive concepts to classify data
Concept	Sensitivity
Relationship(s)	▪ has (data, sensitivity): Relationship shows that data has a sensitivity. ▪ defines (category, sensitivity): A relationship between category that defines sensitivity. ▪ belongs to (sensitivity, variation driver): A relationship between category that belongs to variation driver.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should be able to define data sensitivity based on data category.
Supporting Sources	Attard,et al. (2016), Liu, et al. (2017), Alshammari and Simpson (2018), Gharaibeh et al (2017), Lněnička and Komárková (2019), Ireland. Department of Public Expenditure and Reform. (2018)

Table 17 Concept – Regulations

Design Requirement	DR 4. Provide exclusive concepts to classify data
Concept	Policies/ Regulations
Relationship(s)	▪ belongs to (regulations, variation driver): A relationship between regulations that belongs to variation driver.
Stakeholder(s)	City authorities.
Rationale	A stakeholder should be able to identify which regulations to follow in order to process data to provide services to citizens.
Supporting Sources	Atis (2018), Alshammari and Simpson (2018b), Erl et al. (2016), Elmekki et al. (2019), Plale and Kouper (2017), British Standard (2017), Demchenko et al.

	(2018), Eurocities. (2019), Sutherland and Cook, (2017), Lim, et al. (2018), Ahmad (2022)

Table 18 Concept – Activities

Design Requirement	DR 1. Provide exclusive concepts to manage a data lifecycle
Concept	Activities
Relationship(s)	▪ belongs to (activities, phases): A relationship between activities that belongs to data lifecycle phases.
Stakeholder(s)	City authorities, service providers, service developers.
Rationale	A stakeholder should define activities of a data lifecycle necessary to process data for a specific outcome.
Supporting Sources	Ireland. Department of Public Expenditure and Reform. (2018), Alshammari and Simpson (2018b), Charalabidis (2018), Weber and Kranzlmüller, (2019)

4.3 Phase II: Design

This section defines both the abstract syntax and the concrete syntax that was used to instantiate concepts presented in the previous section.

Abstract Syntax

The abstract syntax in this research is defined using class diagrams from a visual language called UML. To make this abstract syntax easier to understand, concepts are presented as blocks. The concepts are presented in the business and information layers. Figure 20 illustrates the metamodel concepts used to specify data. These concepts derive from Design

Requirement 3 and allow stakeholders to manage data requirements.

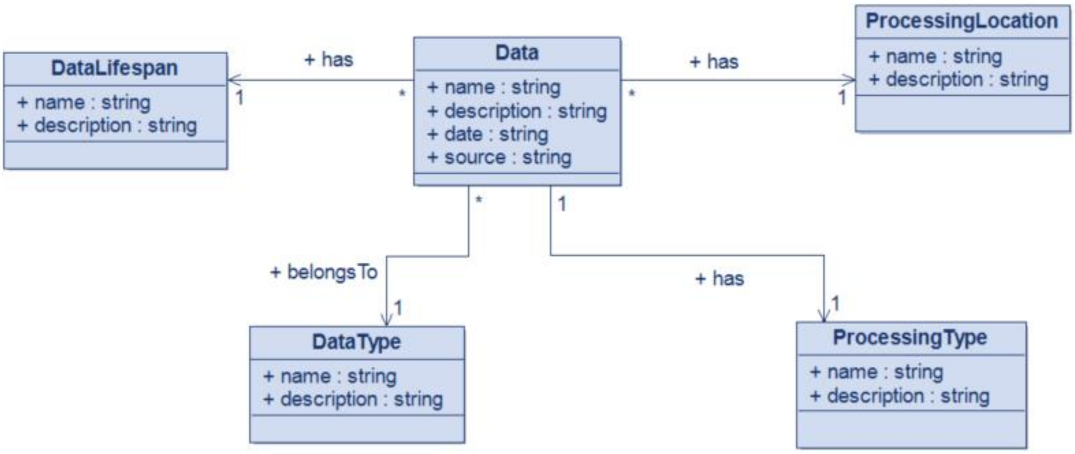

Figure 20 Data Specification

Data Type: Smart cities collect data of different types, which can be unstructured, semi-structured, or structured. This class has two attributes: name and description.

Processing Type: Smart cities also offer several services that have different processing time requirements. This class specifies the type of data processing, which can be batch, real-time, near-time, or streaming.

Processing Location: This class specifies where data will be processed, since sometimes data is processed externally. It is important for stakeholders to know if data will be processed in city councils or in a third-party organisation.

Data Lifespan: This class specifies the lifecycle of data. There are several regulations that cities must follow to process data, so it is important to include this concept in the metamodel to identify when data must be excluded due to a regulation.

Derived from Design Requirement 4, Figure 21 illustrates concepts and their relationships for data classification.

Category: This class organizes data into categories such as biometric, financial, and intellectual property. Categorising data is one of the first steps for classifying it.

Sensitivity: This class defines sensitivity of data, for instance, as public, sensitive, or protected. This concept is directly influenced by the category concept to which data belongs. The category and sensitivity can be defined by a city to reflect the data that will be processed.

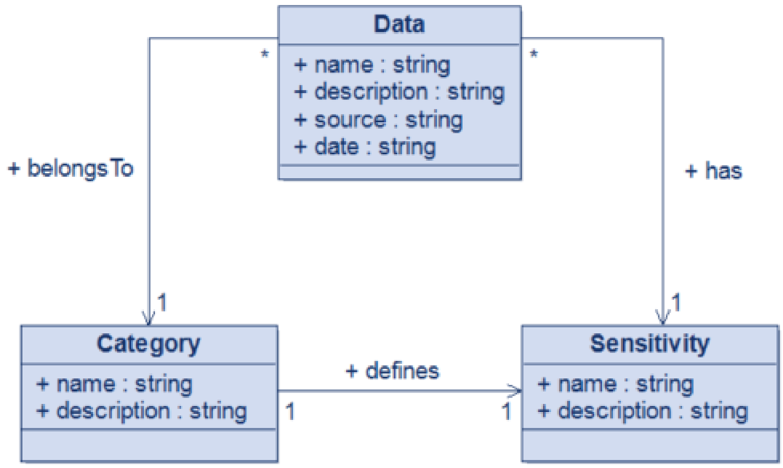

Figure 21 Data Classification

Figure 22 illustrates concepts that typically influence the choice of phases and activities in a data lifecycle. These factors comprise the variation driver that derives from Design Requirement 2. In the smart city domain, the variation driver consists of data lifespan, data type, category, sensitivity, and regulation.

Regulations: This class identifies regulations that cities must comply with in order to process data. The class has three attributes: name, description, and date.

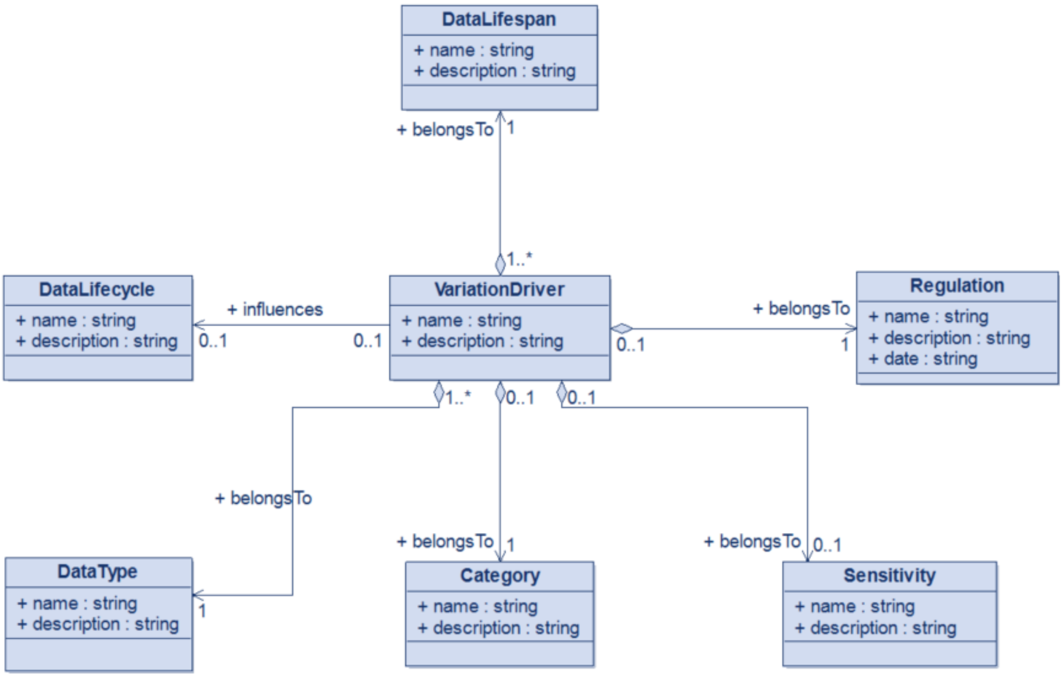

Figure 22 Variation Driver

Deriving from Design Requirement 1, Figure 23 illustrates concepts needed to manage a data lifecycle.

Data Lifecycle: This class identifies a data lifecycle's objective and has three attributes: name, objective, and type.

Phase: This class identifies phases of a data lifecycle that are necessary to reach a certain outcome.

Activity: This class identifies activities that are performed in each phase of a data lifecycle.

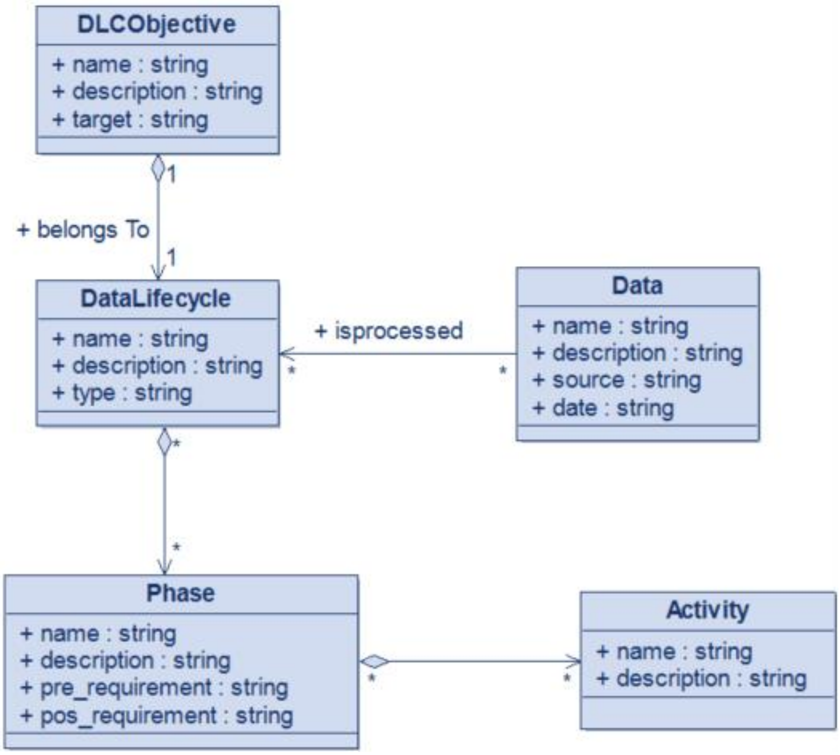

Figure 23 Data Lifecycle Management

Concrete Syntax

Concrete syntax defines the graphical notation used to represent concepts defined by the abstract syntax. Though this research identified new concepts, it used notation elements that already exist in ArchiMate.

4.4 Phase III: Formalise

This phase aims to formalise the metamodel using concepts defined by the identification of requirements. As can be seen in Figure 24, the metamodel at this stage is formalised using UML notation and a UML class diagram. The model illustrates concepts and relationships needed to help stakeholders model variations in data lifecycles. Appendix B illustrates the design requirements of the metamodel.

The new concepts are presented in the information layer, which defines physical and logical data assets and applications used to provide services to citizens. The concepts that populate this layer include data type, data lifetime, processing location, processing type, data lifecycle, phase, activity, category, sensitivity, and variation driver.

The concepts identified in this research are connected by three types of relationships: association, realisation, and aggregation. The meaning of these relationships is described below.

Association: This relationship illustrates a connection between classes; association occurs when the classes need to communicate.

Aggregation: This relationship marks how one class can encompass another class; however, a class (child) cannot exist independently without a class (parent).

Realisation: This relationship represents when a class may have some responsibility for the operations of another class.

Multiplicity – A class diagram also illustrates multiple relationships between elements and defines cardinality, which specifies the number of objects participating in a specific relationship. In addition, role names can be used to define the role of objects.

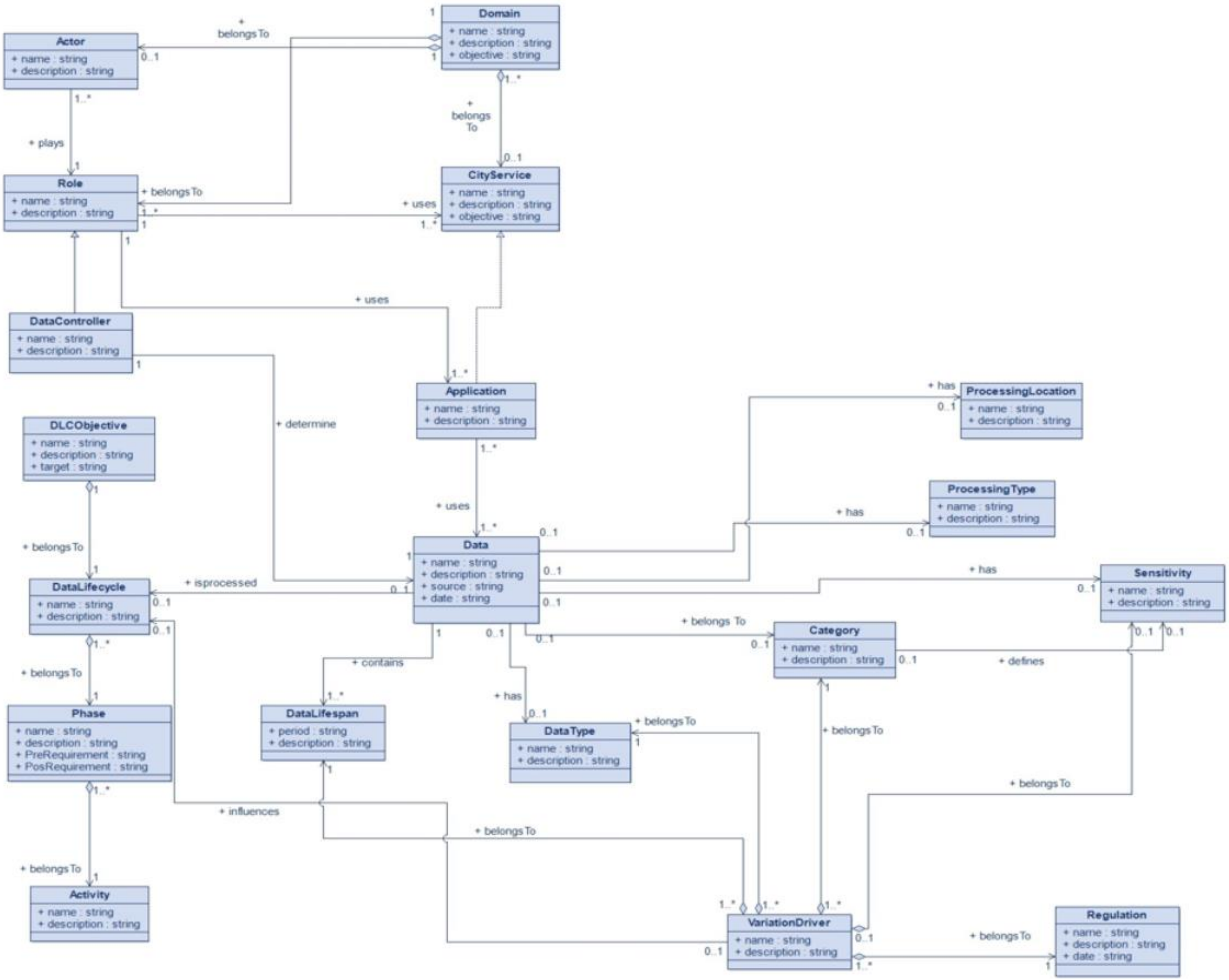

Figure 24 Data Lifecycle Metamodel

4.5 Phase IV: Develop

Design Science aims to build and evaluate an artifact designed to solve organizations' problems (Peffers et al., 2007). This section focuses on developing a metamodel using the proposed concepts identified in Section 4.2.3 using the ArchiMate language. The process followed the Agile Modelling Method Engineering (AMME) approach (Karagiannis, 2015).

4.5.1 ArchiMate Language Overview

A few examples of enterprise architecture modelling languages available include ArchiMate, Architecture of Integrated Information Systems (ARIS) and Multi-Perspective Enterprise Modelling (MEMO) (Lara et al., 2019).

ArchiMate is an open modelling language for Enterprise Architecture used to describe, analyse and visualize different architecture domains and it was developed by the Open Group (The Open Group, 2018).

ArchiMate was chosen as the language to model the proposed concepts for a number of reasons, including its extensive use in the industry and its ability to provide a wide range of elements and domains, enabling the creation of a wide range of models (Rokis and Kirikova, 2023; Atkinson and Kuhne, 2020; Lara et al., 2019). Furthermore, it complies with the ISO/IEC 42010 standard, which serves as a model for architecture description (The Open Group, 2018). Figure 25 depicts the ArchiMate Framework, which has two dimensions: layers and aspects, which are discussed further below.

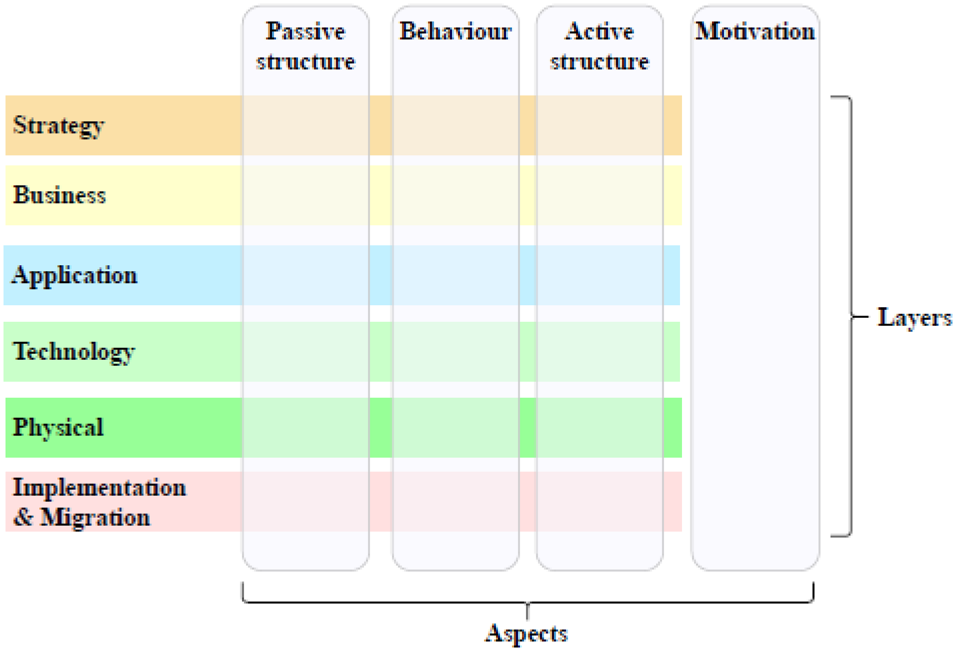

Figure 25 ArchiMate Framework (The Open Group, 2019)

Layers: ArchiMate provides six layers to describe architectures. The language defines concepts and relationships in layers such as strategy, business, application, technology, physical and implementation & migration.

- *strategy layer:* this layer is used to model an organization's strategic direction and decisions.
- *business layer:* this layer focuses on the business aspects, including business processes, actors and products.
- *application layer:* this layers models software applications to support business processes and it provides a bridge between business and technology layers.
- *technology layer:* this layer models the technology infrastructure to support applications.
- *physical layer:* it models the physical infrastructure including servers, storage devices and facilities.

- *implementation & migration layer:* it supports the modelling of implementation and migration of architectures.

Aspects: The aspects are concepts used to represent different perspectives or views of an enterprise architecture. They are categorized in passive, behaviour, active and motivation and are illustrated in Figure 26.

- *passive structure:* it illustrates the concepts on which behaviour is performed such as information and data objects.
- *behaviour:* it focus on behaviour performed by actors. It illustrated the dynamic aspects of an architecture including processes, functions, events and services.
- *active structure:* it focuses on concepts that can perform behaviour including business actors, application components, infrastructure elements.
- *motivation:* it contains concepts used to model motivations or reasons that influence construction or change of an enterprise architecture.

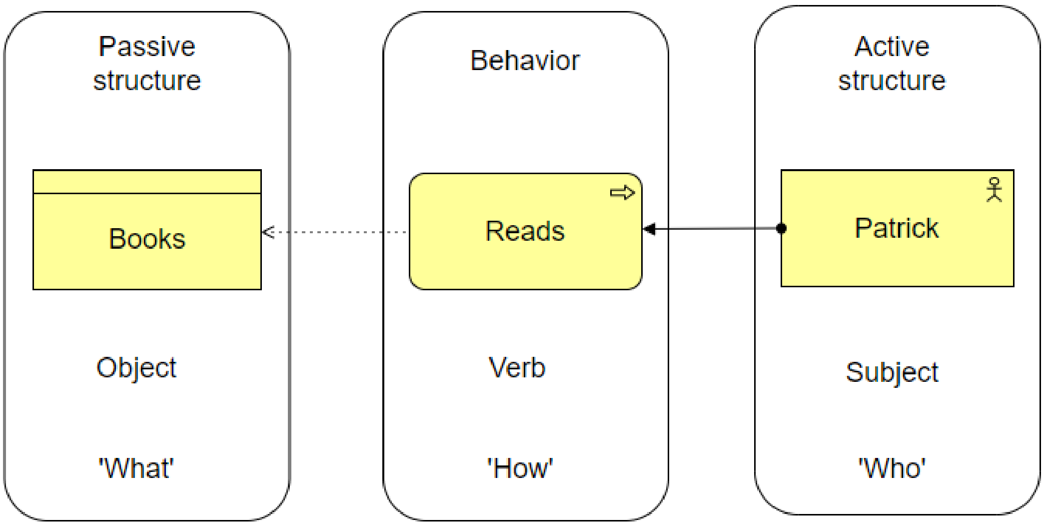

Figure 26 Aspects of ArchiMate language

4.5.2 Metamodel Extension

The new concepts identified in this study (see Section 4.2.3) provide the augmentation of the metamodel, which means the new concepts are from a different problem domain than the original (Atckinson, 2015). This research extends the ArchiMate metamodel using concepts from the smart cities domain.

Mapping concepts

To extend the metamodel, the identified concepts (see Section 4.2.3) must be mapped to ArchiMate's current concepts. ArchiMate 3.1 served as the metamodel extension's foundation. ArchiMate specification was used to help with design decisions for the mapping of business, application, technology and motivation concepts (The Open Group, 2019). Table 19 lists the mapping of data lifecycle metamodel to the ArchiMate concepts. The *Regulation* concept is expressed by the *Requirement* concept of ArchiMate (motivation layer). The ArchiMate metamodel does not include the other concepts.

Data Lifecycle Metamodel		**ArchiMate**	
Layer	**Concept**	**Layer**	**Concept**
Motivation	Regulation	Motivation	Requirement
Information	Sensitivity	Information	New Concept
Information	Category	Information	New Concept
Information	Processing Location	Information	New Concept
Information	Processing Type	Information	New Concept
Information	Data Type	Information	New Concept
Information	Data Lifespan	Information	New Concept
Information	Data Lifecycle Objective	Information	New Concept
Information	Data Lifecycle	Information	New Concept
Information	Phase	Information	New Concept
Information	Activity	Information	New Concept
Information	Variation Driver	Information	New Concept

Table 19 Mapping Concepts

Metamodel Extension

The extended metamodel using the syntax of the ArchiMate language is shown in Figure 27. The relationships between the new concepts are inherited from existing ArchiMate concepts, and the new concepts are organised in the motivation and information layers. The primary ArchiMate concepts are found in the business layer, including business concepts (represented by the colour yellow) and motivation concepts (represented by the colour purple). The application layer's primary ArchiMate concepts are displayed in the information layer and are represented by the colour light blue. The green colour represents the new concepts, which are discussed below.

Information Layer: This layer contains the concepts: *Processing Location, Processing Type, Variation Driver, Regulation, Category, Sensitivity, Data Type, Data Lifespan, Data Lifecycle, Objective, Phase and Activity*. *Processing Location* and *Processing Type* are new concepts and are associated with *Data* concept. *Activity* concept is a aggregation of *Phase* concept, which is a aggregation of *Data Lifecycle* concept. *Data Lifespan* and *Data Type* are new concepts which are associated with *Data* concept and are part of Variation Driver concept. The *Category* concept is a new concept related to *Data* object, it defines the *Sensitivity* concept and is part of the *Variation Driver* concept. The *Sensitivity* concept is related to Data object. The *Requirement* concept from ArchiMate is used to express regulations and policies that a city needs to be in compliance with when processing data. The Driver concept in ArchiMate is used to express the *Variation Driver* and represent the drivers that influence the variations in data lifecycles. *A data controller,* who determines how data will be processed, is represented by the *City Role concept in ArchiMate.*

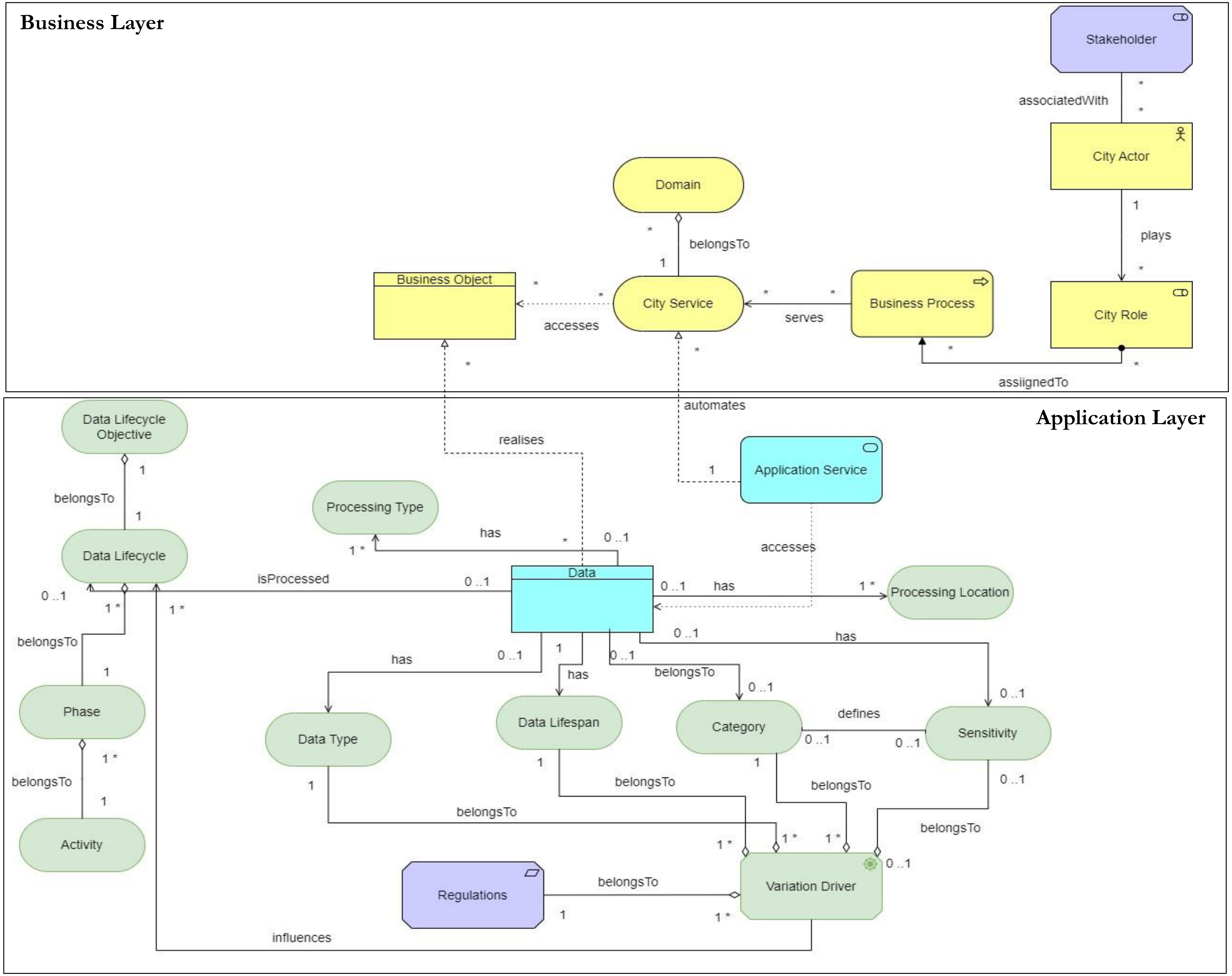

Figure 27 Extended Metamodel

4.6 Summary

This chapter has described three phases for developing a metamodel that can help stakeholders model variations in the data lifecycles of a smart city. The development phases follow AMME, which has six iterative phases. In the first phase, the design requirements are identified, followed by the definition of the smart city domain concepts.

The chapter has also described design requirements and the concepts derived from them, showing relationships, stakeholders, rationale, and supporting sources. Overall, the metamodel represents concepts that are needed to improve the modelling of data lifecycles in the smart city domain.

Chapter 5

5. Demonstration and Evaluation

This chapter addresses RQ3, which is to demonstrate and evaluate the metamodel developed to model data lifecycle variations. Section 5.1 of the chapter provides an overview of the metamodel's demonstration and evaluation. Section 5.2 demonstrates how the developed artifact may be used and applied. Case studies were undertaken during this phase of research to demonstrate the use of real-world scenarios. Section 5.3 describes the evaluation process, along with evaluation criteria and findings; it also includes an assessment involving case studies and validation of the metamodel by smart city practitioners. The validity and reliability of these research findings are strengthened by using the triangulation method (see Chapter 3). Section 5.4 summarises the chapter.

5.1 Demonstration and Evaluation Overview

This section presents an overview of the demonstration and evaluation phases of the study. Demonstration and evaluation is one of the DSRM phases, as previously presented in Chapter 3 and illustrated in Figure 15. The goal of this phase is to understand the limitations of modelling data lifecycles and whether new concepts need to be identified and validated. The demonstration phase shows the use of the metamodel in case studies, as described below.

Cork Case Study

The Cork case study was conducted during the demonstration and evaluation phases of DSRM. For this case study, EA models were made using ArchiMate in order to get feedback from practitioners and evaluate drawbacks in modelling data lifecycles.

Design requirements, along with concepts and relationships, were also evaluated by practitioners by asking their relevance to the smart city domain thus validating the design metamodel. The metamodel was evaluated by practitioners using the evaluation criteria presented in Section 5.3. Practitioners assessed design requirements, concepts, and relationships based on their applicability to the domain, validating the design metamodel.

Limerick Case Study

The Limerick case study was also conducted during the demonstration and evaluation phase of DSRM. For this case study, design requirements, concepts, and relationships were assessed by practitioners in order to obtain their feedback on their relevance for the smart city domain and to validate the design of the metamodel. The metamodel was evaluated by practitioners using the evaluation criteria presented in Section 5.3.

5.1.1 Phase V: Deployment and Validation

This section describes how the validation phase of the metamodel is conducted. In this phase, the artifact is deployed and its use is evaluated. In this research, the metamodel was evaluated by practitioners and their feedback were taking into consideration for the next iteration. Figure 28 illustrates the phases of AMME that were followed to evaluate the metamodel. The next sections present the demonstration and evaluation phases of this research.

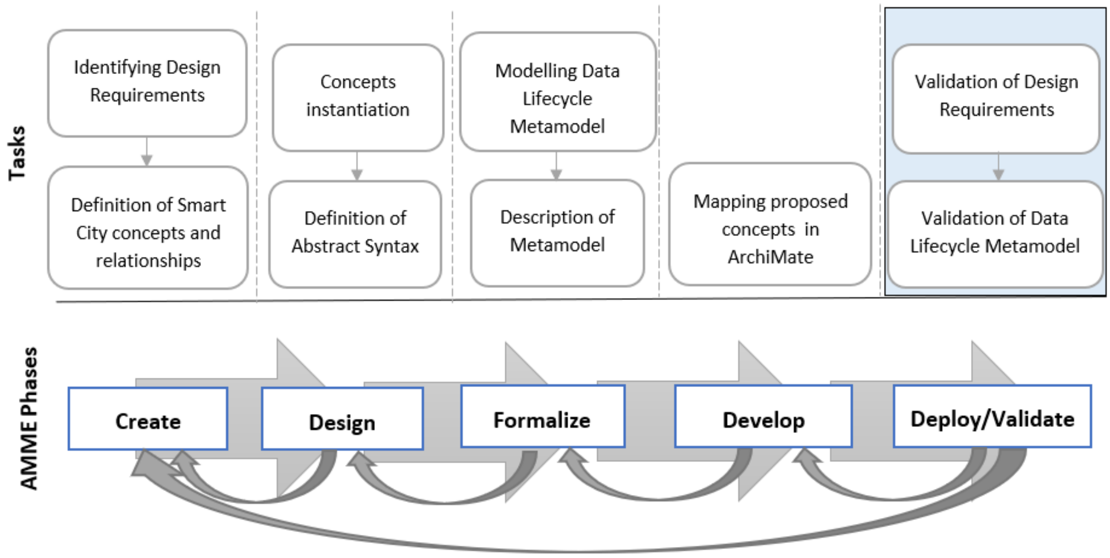

Figure 28 Metamodel Development Method (AMME) Validation Phase

5.2 Demonstration

The demonstration is a very important phase where it can validate the practical relevance and effectiveness of a developed artifact. It is a iterative process, where a researcher observes how well the artifact achieves its main purpose. In this research, two case studies were conducted to verify whether the study was successful in identifying the concepts required to represent the variations in data life cycles required in data management and thus formalise the concepts and their relationships in a metamodel. The next two sections respectively present the case studies that were conducted in the cities of Cork and Limerick in Ireland.

5.2.1 Cork Case Study

Cork City Overview

Cork is a city in County Cork located in the southwest of Ireland. Based on preliminary results of the last census, Cork's population is over 222,000, which makes Cork the third largest city by population in Ireland. The city covers an area of 174 km^2 with a population density of 1,188/km^2. Cork has launched Cork Smart Gateway, an initiative to transform Cork into a smart, sustainable, and inclusive place with the assistance of academic partners, and its focus is technology, data, and digital tools. The city also has a digital strategy with actions that focus on seven dimensions: citizen participation, support services, digital skills, open data, climate action, infrastructure, governance and leadership.

CCTV City Service

A Closed Circuit Television Systems service (CCTV Systems) was selected by senior manager from Cork City Council since it processes sensitive information and thus must meet different processing requirements. The CCTV city service is used in cities to create a safer environment, assist with better traffic management, and help police with criminal investigations.

Cork City has several cameras installed across the city and county; however, their exact locations are kept secret to prevent vandalism, as they are used to combat illegal dumping too. Currently, the Cork City Council controls two types of CCTV: private and public.

- The private CCTV System is used in places where the public is not allowed to enter, such as fire stations, council buildings, and operational depots.
- The public CCTV System is used in public areas that the public has the right to access, such as streets, parks, cemeteries, roadways, and bridges.

The Cork City Council follows a CCTV and audio recording policy since the CCTV system can infringe on the rights of individuals by capturing personal data like their images. The Data Protection Acts of 1988 to 2018 establish a maximum of 28 days for retaining data from CCTV, except when the images are used in a criminal investigation or court proceedings; otherwise, the recordings must be erased, and all recordings must be placed in secure places with restricted access.

Data Collection and Data Analysis

According to Yin (2014), a significant advantage to collecting data during a case study is the ability to acquire data from multiple sources. The research in the current study is based on three main data sources: meetings, semi-structured interviews, and secondary data.

An ethical approval was granted to conduct interviews, and interview subjects additionally signed a consent form (see Appendix A). Four semi-structured interviews with senior managers from the Cork City Council were conducted. The interviews lasted from 30 to 45 minutes and were conducted online. With the consent of the interview subjects, the interviews were taped, transcribed, and saved for data analysis. The questions asked during the interviews were related to the data management of city services.

Secondary data for this study includes the official city council website and reports. The CCTV system was chosen because of the importance of this service and the sensitivity of the data collected. During analysis of this data, concepts found in the literature were coded in conjunction with the following: case studies related to the main findings of this study, design requirements, and concepts and their relationships that support modelling of variations in data lifecycles.

Instantiation of Metamodel

In this section, the researcher developed architecture models of Cork City's CCTV system that accord with the metamodel proposed by this research. The models were developed following a top-down approach, starting from the business layer and ending with the information layer. The models developed for the Cork City case study are described below.

Stakeholders View

Figure 29 illustrates the stakeholders view, which represents a group of stakeholders for CCTV service in Cork. Stakeholders are divided into internal and external groups. Internal stakeholders refer to different city council departments (the smart city department, planning department, and security department) that have access to internal data. External stakeholders refer to stakeholders outside the city council, such as police, data subjects, and other third parties. These stakeholders can only access CCTV recordings after requesting access to them through a data protection officer.

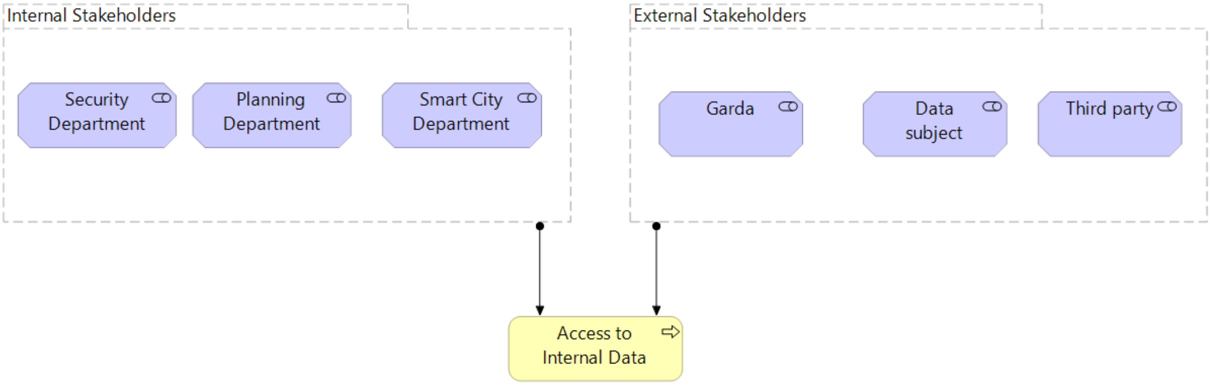

Figure 29 Cork Case Study - Stakeholders View

Motivation and Requirements View

Figure 30 presents the motivation and requirements view for CCTV service in Cork. It illustrates the requirements of the application, data, and technology layers. It also demonstrates how the goal "to secure public order and safety in public places" is realised by the outcome "to ensure that CCTV data will be used for a specific purpose". This outcome – which is realised by specific requirements at the application, data, and technology layers – aligns with the CCTV and Audio Recording Policy and Procedure issued by Cork City and with the Guidance for Data Controllers issued by the Data Protection Commission of Ireland.

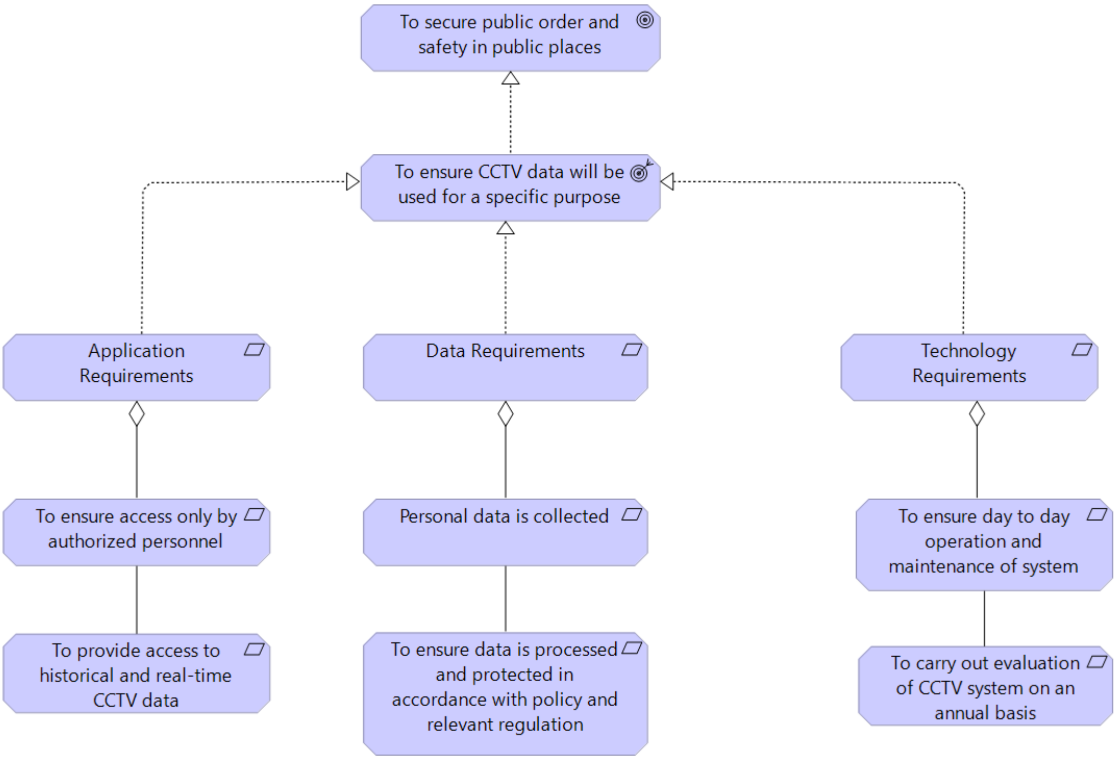

Figure 30 Cork Case Study – Motivation and Requirements View

Information Structure View

Figure 31 illustrates the structure of information used by the service. The view shows information at the data level. The figure illustrates the data structure of this service, which contains the date, time, location where a camera is located, and images at this specific location. The purpose of this view is to show what data is collected during the service. This view provides information that will be needed to decide the category and sensitivity of data and, consequently, how they will be handled.

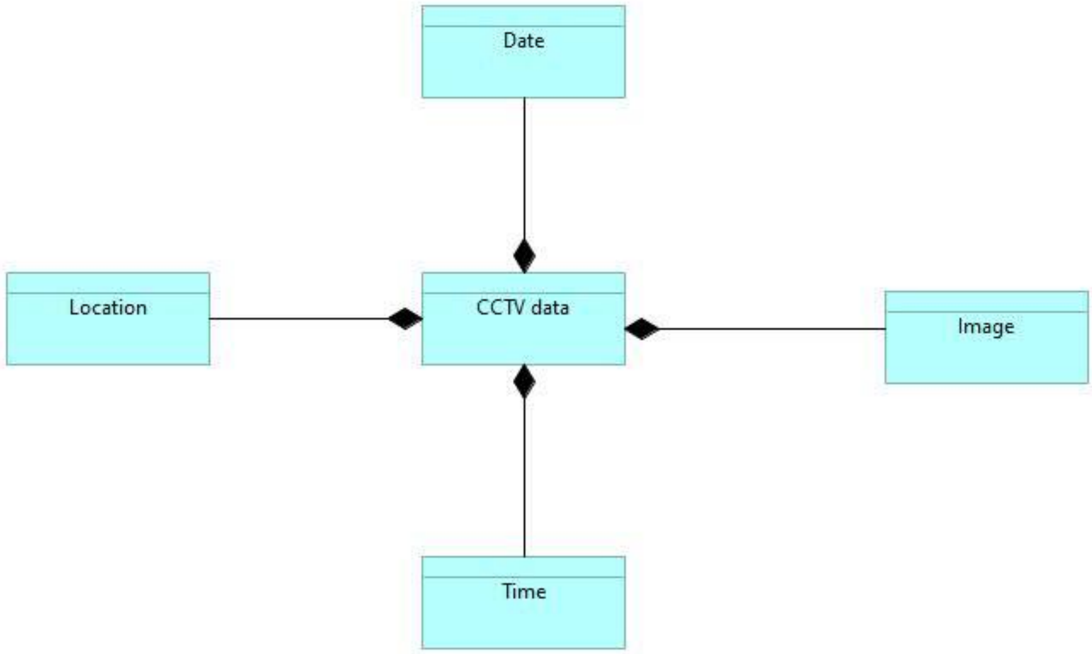

Figure 31 Cork Case Study – Information Structure View

Data Requirements View

Figure 32 illustrates the data requirements view, which represents concepts for managing data requirements that need to be considered for CCTV service. The category of this data is classified as personal data and the sensitivity as sensitive. Moreover, the model informs the processing location, processing type, data lifespan, data type, and regulation that this data must comply with. This model assists in defining data processing requirements. In this service, regulations require that data be kept for only 28 days, except for data used for police investigations. Because data is classified as sensitive, there is also a requirement to protect its processing and storage.

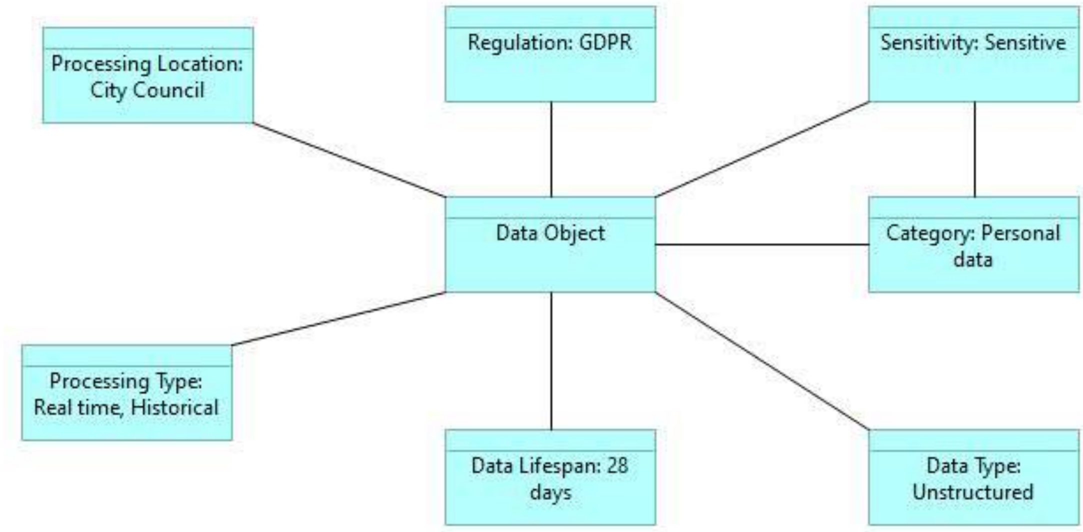

Figure 32 Cork Case Study – Data Requirements View

Variation Driver View

Figure 33 illustrates the variation driver view, which represents the drivers that need to be considered for defining the data lifecycle phases and activities of CCTV service in Cork City (see Section 2.2.3). Each city service has its own requirements that influence drivers. In the CCTV service, data is defined as sensitive, so it cannot be shared. Defining drivers is important for data management, since it defines how data will be processed, stored, and protected.

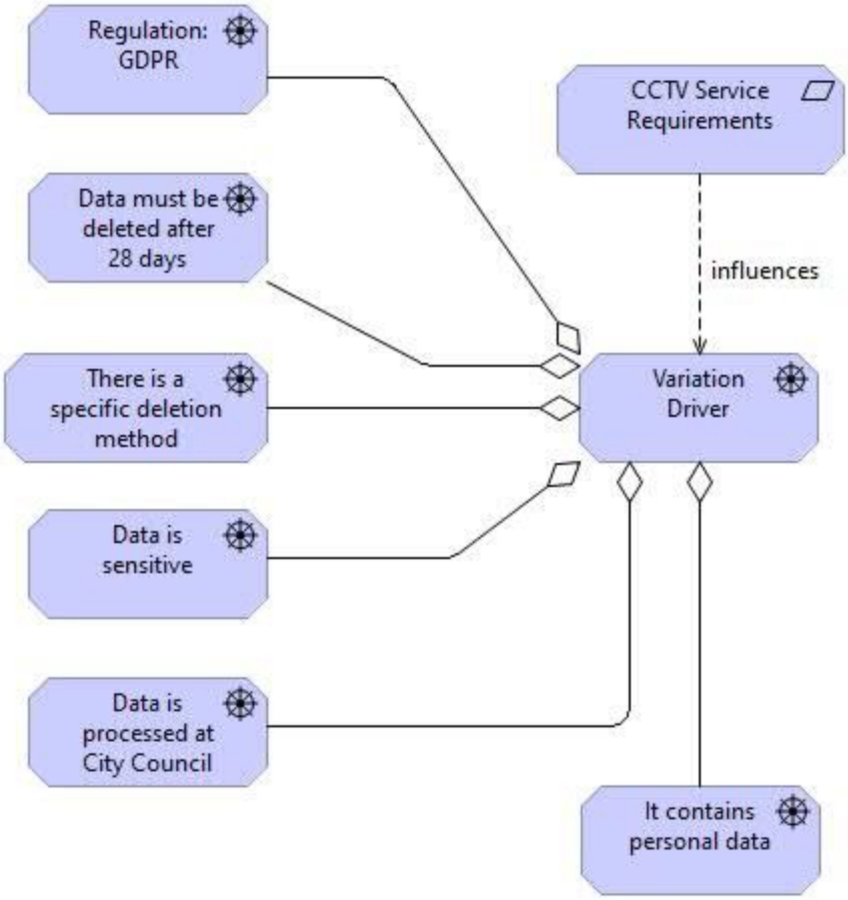

Figure 33 Cork Case Study – Variation Driver View

Conclusion

The case study demonstrates the applicability of proposed concepts that support the variations in data lifecycles. The usage of the metamodel shows that these concepts are relevant to support the representation of data lifecycle variations.

A modelling technique proposed by TOGAF was followed from top to bottom, starting with the motivation layer, extending to the business layer, and ending with the information layer. Table 20 presents the list of proposed concepts in the metamodel and its application in the Cork case study. The symbol √ indicates the necessity of a concept in the models designed during the case study.

The manager at Cork City Council uses the models to communicate the solutions design to other stakeholders (e.g. digital officer, head of operations). The case study demonstrated to city managers that the models can help them in planning for changes in city services prior to developing new applications or solutions. And we can assert that the concepts proposed by this study take into account GDPR requirements.

A senior manager at Cork City Council provided a feedback stating that the concepts offer a structured approach to the design and transformation of public services. Additionally, the models can help make services more transparent to citizens, as last year, a different municipality was fine €110,000 for misusing data from the CCTV service when it was revealed that the board did not respect the requirements of the GDPR.

Table 20 Cork Case Study

Layer	Data Lifecycle Metamodel	Cork Case Study
Motivation	Variation Driver	√
Information	Data Lifespan	√
Information	Sensitivity	√
Information	Category	√
Information	Processing Type	√
Information	Data Type	√
Information	Data Lifecycle Phase	√
Information	Data Lifecycle Activity	√
Motivation	Regulation	√
Information	Processing Location	√

5.2.2 Limerick Case Study

Limerick City Overview

Limerick is a city in western Ireland, located in County Limerick, with an area of 59.2 km^2. Limerick's population of 205,444 makes it the fourth most populous city in Ireland. Limerick has been chosen to be one of the two lighthouse cities in the +CityXchange Smart City project, which aims to create a sustainable urban ecosystem. Limerick also has a strategy to use digital technologies to improve the quality of life of its citizens by developing and improving services for them.

Footfall Counter Service

A footfall counter is a service that uses sensors to capture the number of pedestrians in a certain place in a city or rural area. The data collected by this service, which shows how specific areas are utilised, is used mainly for tourism, events, and retail development. Limerick installed its first counters in 2016 on the main intersection of O'Connell Street and Sarsfield/William Street. Figure 34 depicts an image of the high-precision pedestrian counter, which uses thermal imaging.

Open data is critical to the development and operation of smart cities for a variety of reasons, including promoting transparency, citizen participation, and engagement (Gao et al., 2023). Due to the importance of open data, this study chose a city service that provides open data to its citizens to analyse data management requirements in this scenario too. Therefore the footfall counter service was chosen by Limerick city managers since it does not process sensitive information, which distinguishes it from the CCTV service in the Cork case study.

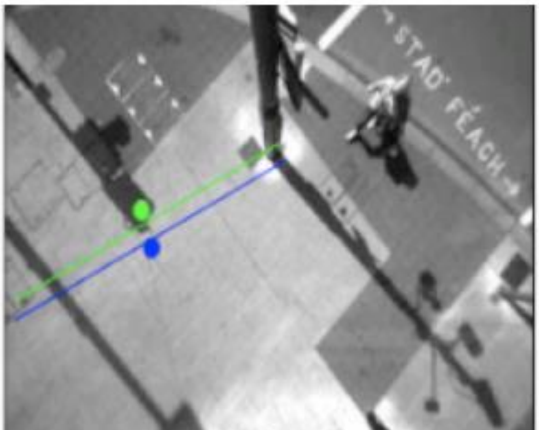

Figure 34 Footfall Counter – Limerick

Figure 35 presents how data is collected in the footfall counter service. The excel file, which is available on the city council website, contains the date, time, and number of IN and OUT pedestrians in zone 1.

	A	B	C
1	count_datetime	zone_1_in	zone_1_out
2	20/04/2018 00:00	13	3
3	20/04/2018 00:15	14	7
4	20/04/2018 00:30	16	4
5	20/04/2018 00:45	5	3
6	20/04/2018 01:00	5	5
7	20/04/2018 01:15	12	3
8	20/04/2018 01:30	9	5
9	20/04/2018 01:45	10	3
10	20/04/2018 02:00	3	8
11	20/04/2018 02:15	12	1
12	20/04/2018 02:30	51	8

Figure 35 Footfall Counter data

Data Collection and Data Analysis

This case study is based on three main data sources: meetings, semi-structured interviews, and secondary data. Two semi-structured interviews were conducted with a senior manager from Limerick City and County Council. The interviews ranged from 30 to 45 minutes, and they were conducted and recorded online due to Covid-19. A transcript of each interview was created and stored for data analysis. The interview questions addressed each stage of the service design. Secondary data in the case study includes the official city council website and reports.

Analysis involved coding and categorizing the data using the inductive approach proposed by Thomas (2006). Themes and categories were identified in the interview transcripts to help condense the raw text and align the interview findings with the research objectives of this research. NVivo software was used to carry out data analysis. The footfall counter service was chosen because provides an opportunity to show variation in data lifecycle modelling.

Instantiation of Metamodel

This section describes how architecture models of the footfall counter service in Limerick were developed in compliance with the metamodel proposed by this research. The models follow a top-down approach, extending from the motivation layer to the information layer. The models developed for the Limerick City case study are described below.

Stakeholders View

The city council aims to provide footfall counter service to a wide number of stakeholders in Limerick City. Figure 36 illustrates the group of stakeholders for this service, both internal and external. Internal stakeholders refer to different departments of the city council (tourism strategy, smart travel, and economic development) that have access to internal data. External stakeholders refer to stakeholders outside the city council, such as citizens and retailers, who have access to public data made available on the city council website.

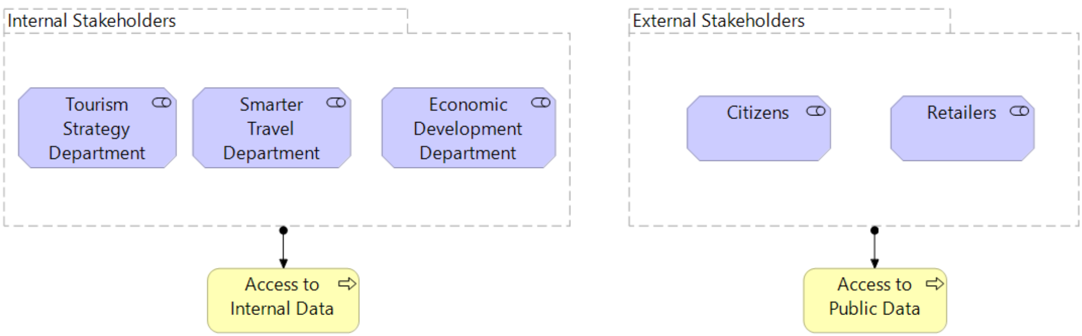

Figure 36 Limerick Case Study – Stakeholders View

Motivation and Requirements View

Figure 37 depicts the motivation and requirements view for the footfall counter service in Limerick. It illustrates the requirements of the Application, Data, and Technology layers. It also demonstrates how the goal to "assist decision making to improve design and development of the city" is realised by the outcome to "provide footfall counter information". This outcome aligns with Ireland's Digital Strategy to "collect once, use it several times", which follows Estonia's principle of not duplicating data. This principle is realised by specific requirements in the application, data, and technology layers.

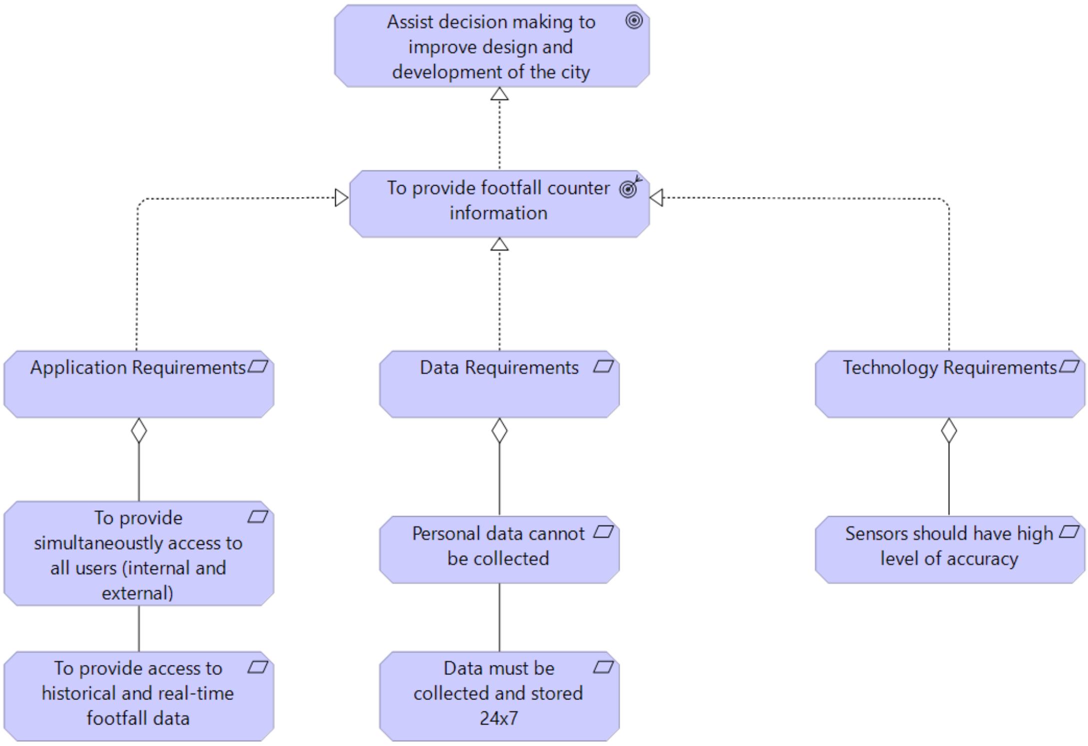

Figure 37 Limerick Case Study – Motivation and Requirements View

Information Structure View

Figure 38 illustrates the structure of the information used by the footfall counter service. The view shows information at the data level. The figure illustrates the data structure of the service, which contains the date, time, zone where a sensor is located, and number of pedestrians IN and OUT at a specific location. The purpose of this view is to show what data is collected during the service, which is also shown in Figure 35. This view provides information that will be needed to decide the category and sensitivity of data and, consequently, how each will be handled.

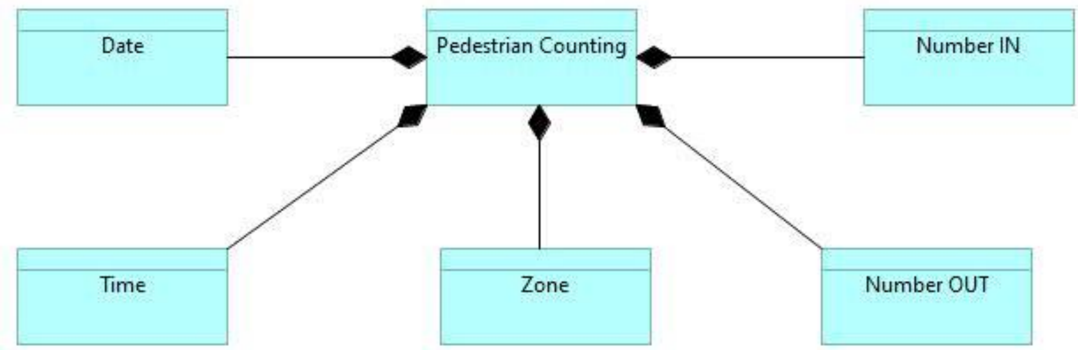

Figure 38 Limerick Case Study – Information Structure View

Data Requirements View

Figure 39 illustrates the data requirement view, which represents concepts for managing data requirements that need to be considered. The category of this data is classified as city data and the sensitivity as public, and the model foregrounds the processing location, processing type, data lifespan, data type, and regulations that data must comply with. This model assists in defining data processing requirements; for instance, the data lifespan is defined as 10 years, so after 10 years the data will be deleted. Because data is classified as public, there is no need to specify a deletion method to prevent data recovery.

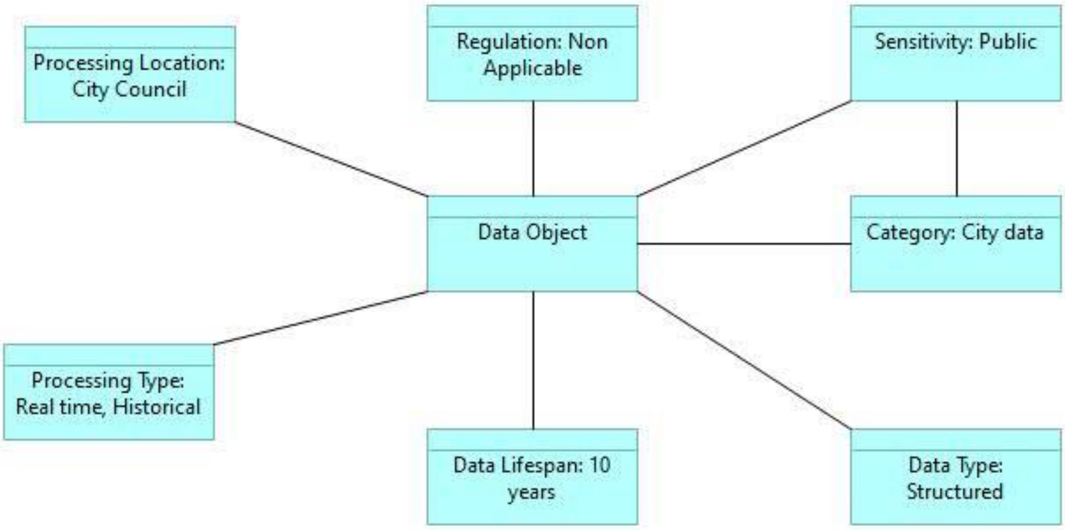

Figure 39 Limerick Case Study – Data Requirements View

Variation Driver View

Figure 40 illustrates the variation driver view, which represents the drivers that need to be considered for defining data lifecycle phases and activities. Each city service has its own requirements that influence drivers. In the Footfall Counter Service, data is defined as public, so it does not need to be made anonymous and can be fully published as shown in Figure 35. Defining drivers is important for data management, since it defines how data will be handled.

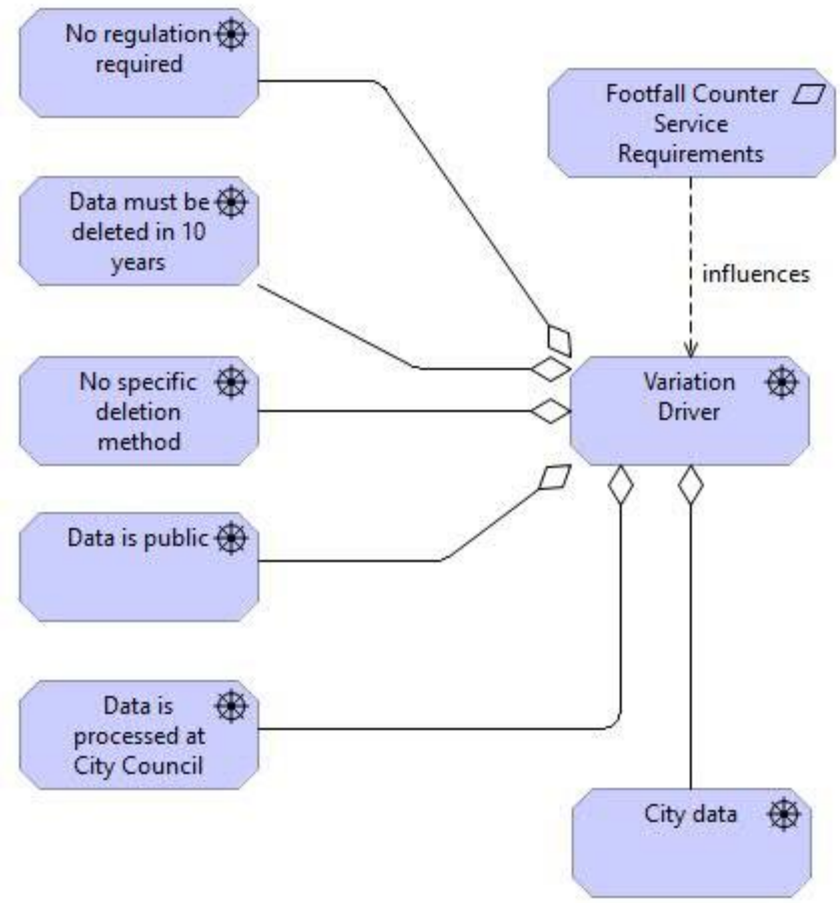

Figure 40 Limerick Case Study – Variation Driver View

Data Lifecycle View

Figure 41 presents the data lifecycle view, which represents the phases of a data lifecycle. The figure presents a structural overview of the target architecture, which entails details concepts that are specified in the motivation, business, and information layers. The motivation layer presents the drivers that influence data lifecycle phases and activities decisions. The business layer presents internal and external users of the service. And the information layer presents application interfaces offered by the service to provide access to data; it also represents the data lifecycle phases necessary to process data.

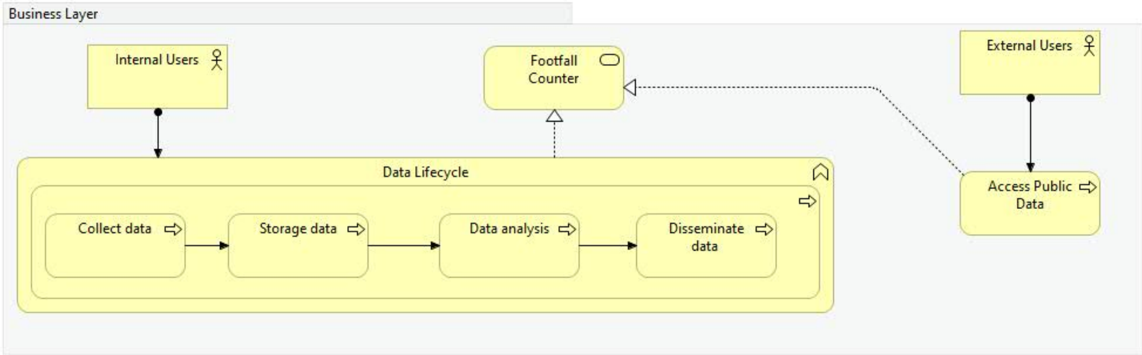

Figure 41 Limerick Case Study – Variation Driver View

Conclusion

The Limerick case study was conducted to show the applicability of concepts identified in this research, which aims to model variations that occur in data lifecycles. The case study allowed the researcher to replicate findings from the Cork case study and to validate them in a different city by modelling a Footfall Counter Service. A modelling technique proposed by TOGAF was followed from top to bottom, sequentially addressing the motivation, business, and information layers, in that order.

Table 21 presents the list of concepts in the proposed metamodel and its application in the Limerick case study. The √ indicates the necessity of a concept in models designed during the case study.

Table 21 Limerick Case Study

Layer	Data Lifecycle Metamodel	Limerick Case Study
Motivation	Variation Driver	√
Information	Data Lifespan	√
Information	Sensitivity	√
Information	Category	√
Information	Processing Type	√
Information	Data Type	√
Information	Data Lifecycle Phase	√
Information	Data Lifecycle Activity	√
Motivation	Regulation	√
Information	Processing Location	√

The purpose of the Limerick case study was to investigate the use of the proposed concepts in a service that processes non-confidential information. Footfall data is available online, and this case study demonstrated that neither TOGAF nor ArchiMate provides the concepts necessary for modelling data lifecycle variations. As described in Section (1.1.1), data lifecycles vary to meet different data processing requirements. For example, footfall data is public, so there is no specific method for deleting it when the data lifespan is reached. The proposing concepts and their relationships are fundamental for representing data lifecycles and their variations, for providing stakeholders with a holistic view of the data processing related to various services offered to citizens, for helping in decision-making, and for verifying whether city councils are complying with regulations and data management requirements.

Table 22 lists the variations points and options for both case studies. Variability in an artefact is described using variations points and options. Variation points, as defined in Section 1.1.2, represent the locations of variations in an artifact's feature or process. Variation options, on the other hand, represent how an artefact varies. Variation points and options are utilised in this study to show variability in data lifecycles.

Table 22 Variation Points and Variation Options on both case studies

Variation Points	**Variation Options**	
	CCTV Service - Cork	Footfall Counter - Limerick
Phases	- Collect - Storage - Preparation - Analysis - Usage - Delete	- Collect - Storage - Preparation - Analysis - Sharing - Delete
Activities	- Specific deletion method - Manage data access views and permissions - Data security management	Make data available at city council website
Regulation	GDPR	Non applicable
Data Lifespan	28 days	10 years
Category	Personal data	City data
Sensitivity	Sensitive	Public
Processing type	Real-time, historical	Real-time, historical
Data Type	Unstructured	Structured

5.2.3 Cross-Case Analysis

After conducting the two case studies, the researcher used cross-case analysis to draw out similarities and differences between the cases in order to acquire a deeper understanding of issues across different contexts (Yin, 2014). The case studies were conducted in Cork and Limerick, cities in Ireland, and the models produced for the two cities had to meet different data management requirements, since they involved managing data of different sensitivity. The two case studies allowed the researcher to analyze different scenarios and the different requirements needed in data processing to offer services to citizens. Table 23 presents the data sources used within the case studies. In this study, data was collected to model the variation of data lifecycles related to public city services.

Table 23 Cross-Case Analysis - Data Sources

Case Study	Data Source
Cork City	1. Main data ▪ Meetings ▪ Semi-structured interviews 2. Secondary data ▪ City council website ▪ CCTV & Audio Recording Policy and Procedure ▪ Cork Digital Strategy ▪ CCTV Guidance for Data Controllers ▪ Public Service Data Strategy
Limerick City	1. Main data 2. Semi-structured interviews 3. Secondary data ▪ City council website ▪ Public Service Data Strategy ▪ Limerick Digital Strategy

Data triangulation was used to compare and contrast the findings across data sources for each case study. The findings of this study, including the proposed design requirements and the metamodel, were analysed across the two case studies (see Chapter 5). Primary stakeholders were asked about the usability and applicability of the design requirements and the metamodel. The cross-case analysis is presented below and the data sources that were used to compare the main findings across case studies are shown by the arrows in Figures 42 to 44.

Design Requirements

Figure 42 illustrates the design requirements and data sources used to investigate the findings across the two cases. Four design requirements were identified (see Section 4.2.1) for modelling variations in data lifecycles. For instance, Design Requirement 2 which is required to provide exclusive concepts to manage variability in data lifecycles was validated with the Public Service Data Strategy of the Government of Ireland, which states that "multiple versions of a data lifecycle exist with variations in practices across different business domains".

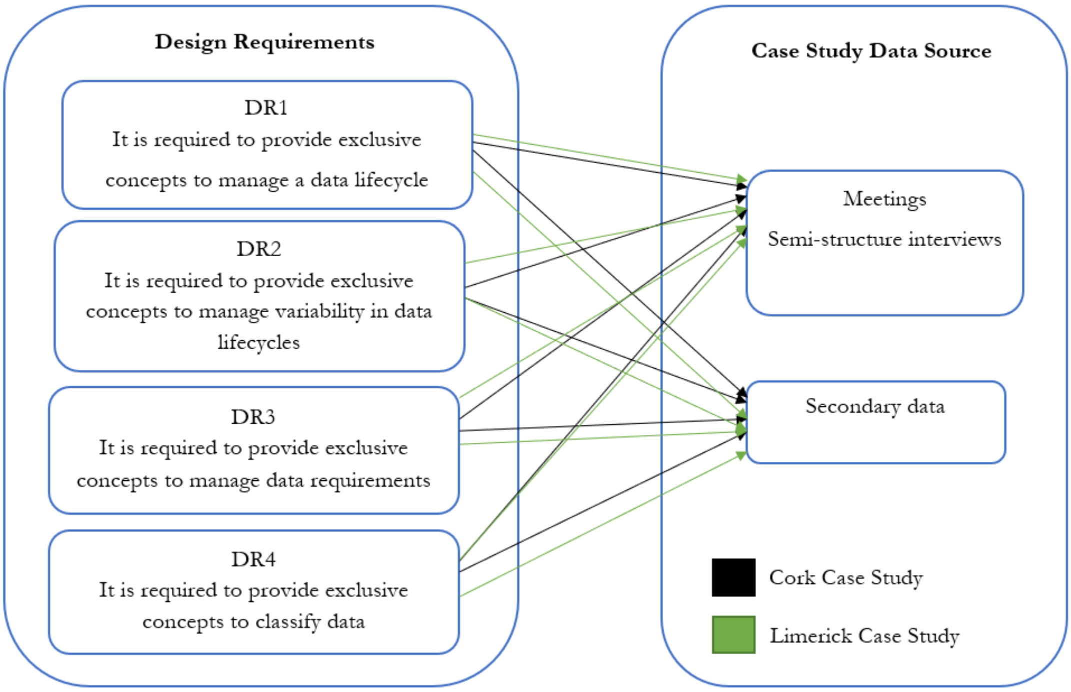

Figure 42 Cross case Analysis – Design Requirements

Motivation Concepts

Figure 43 illustrates motivation concepts and the data sources used to compare outcomes from the case studies. Both cities reported that they do not delete data as a precaution because it will be needed in the future. Therefore, the researcher created a connection between regulation and data through the variation driver, as data processing must comply with regulations and creates guidelines for stakeholder access, methods for deleting data, and storage times for data. The connection was created using an association in the metamodel. Associations are used to represent relationships between classes.

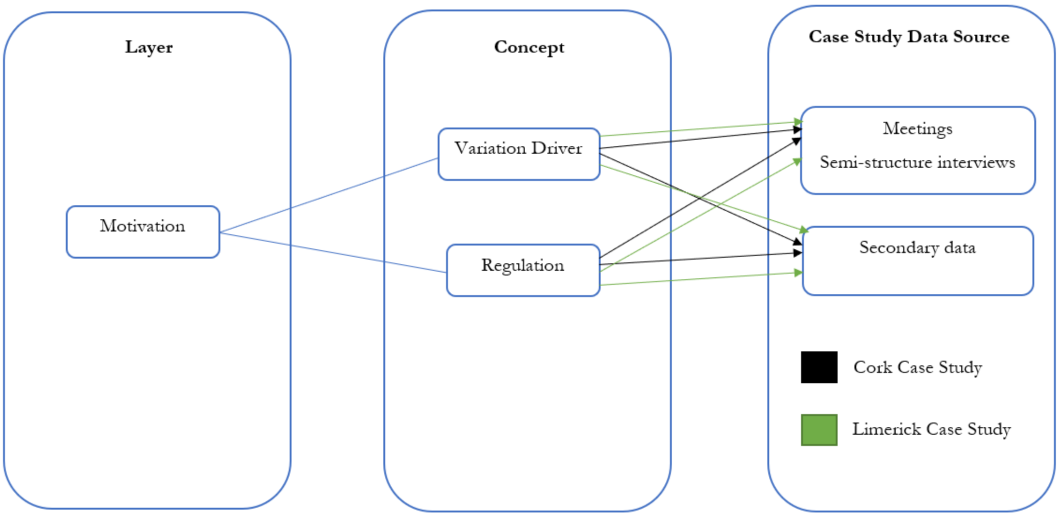

Figure 43 Cross-Case Analysis – Motivation Concepts

Information Concepts

Figure 44 presents the information concepts and data sources used to compare the outcomes of the two case studies. Cork conducted a review of its CCTV service that addressed data collection and access. The objective was to ensure that the collected data was being used in the way that was proposed to citizens. Among the six information concepts, the "objective" concept was created to specify the purpose of a specific data lifecycle. Other information concepts include "category" and "sensitivity", which mainly identify which data is sensitive and needs different handling. Both city councils understood how important these concepts are for processing data from services that have different data sensitivity and requirements.

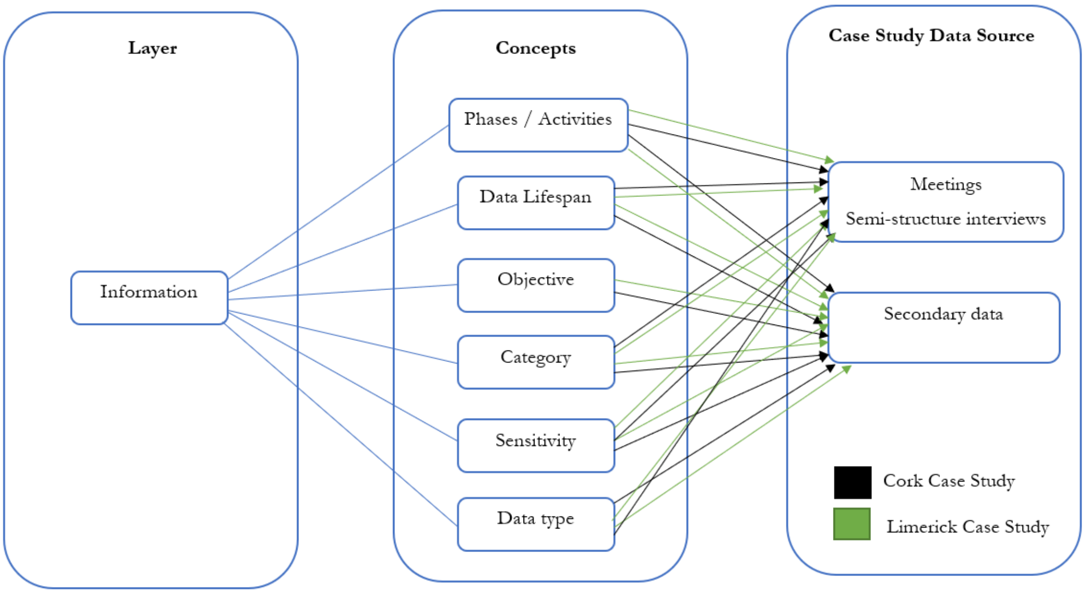

Figure 44 Limerick Cork Case Study – Variation Driver View

5.3 Evaluation

Design Science Research (DSR) consists of two main activities: build and evaluate (Mark and Smith, 1995). As a crucial DSR activity, evaluation must be conducted with rigour to ensure the utility, quality, and efficacy of a designed artifact (Hevner and Chatterjee, 2010). Therefore, evaluation of an artifact must be conducted while it is being built (ex-ante) as well as after it has been developed (ex-post) (Pries-Heje et al., 2008). Section 5.2 demonstrates how the proposed metamodel applies to real-life case studies in two cities in Ireland. Table 24 outlines the outputs of DSR activities in this research and what was evaluated during its ex-ante and ex-post phases. The ex-ante and ex-post evaluations are detailed in the sections that follow.

Table 24 Ex-ante and Ex-post Evaluation

RQs	RQ1	RQ2	RQ3
DSR Activities	Problem and Motivation, Define Objective	Meta design, Artificial Evaluation Design Practice	Demonstration and Naturalistic Evaluation
Outputs	Problem definition, Hypothesis, RQs based on literature review	Identification of design requirements, data lifecycles elements based on systematic literature review and semi structured interviews	Metamodel instantiation
Evaluation	Evaluation of relevance of the problem based on meetings with practitioners	Evaluation of design requirements, concepts and their relationships based on discussion with experts	Evaluation of metamodel concepts and their relationships based on case studies
	Ex Ante		**Ex Post**

5.3.1 Ex-Ante Evaluation

Ex-ante evaluation is performed to validate an artifact's design (Sonnenberg and Broke, 2012). To conduct the ex-ante evaluation, the researcher followed the evaluation activities proposed by Sonnenberg and Brocke's (2012), as shown in Figure 45. The authors divide the ex-ante evaluation into two phases: evaluation 1 and evaluation 2.

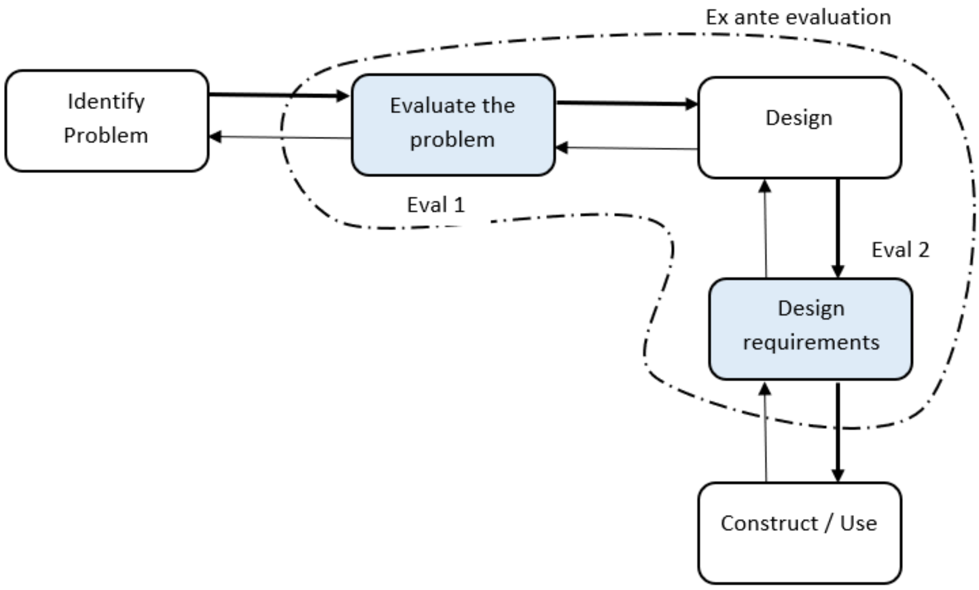

Figure 45 Ex-ante evaluation activities from Sonnenberg and Broke (2012)

Evaluation 1 occurs after problem identification to ensure that the identified problem is relevant to practice and adds novelty to research. The problem statement and objectives are inputs for this evaluation phase. In this study, a literature review was first undertaken to identify gaps in the domain and to compare research findings with existing knowledge from previous studies. In order to validate the results from the literature review, the researcher then held meetings with practitioners from the ISO standard group. In sum, evaluation 1 supported the design of the researcher's artifact by validating the novelty and importance of the research problem through a literature review performed in conjunction with expert meetings.

Evaluation 2 is carried out to ensure that the artifact's design satisfies the study objectives. Design objectives are this phase's input. The elements of a data lifecycle were identified, and they served as a starting point for defining the design requirements and concepts for the artifact that was subsequently created. A literature review and meetings with experts were also used to determine whether the identified concepts and relationships represent all the statements that the final artefact must have. The inputs and outputs of each ex-ante evaluation step are shown in Table 25 along with the criteria and methods used.

Table 25 Ex-Ante Evaluation

Activity	Input	Output	Criteria	Methods
Evaluation 1	- Problem statement - Objectives	- Justified problem statement and objectives	- Novelty - Importance	- Literature Review - Meetings with experts
Evaluation 2	- Design objectives	- Validated design requirements, concepts, relationships	- Completeness	- Literature Review - Meetings with experts

5.3.2 Ex-Post Evaluation

An ex-post evaluation is used to assess an artefact after it has been created and used. In this study, the ex-post evaluation focussed on the metamodel that the researcher had developed, and it was guided by Research Question 3: “How do we evaluate the proposed concepts to support the modelling of data lifecycle variations in enterprise architectures?” The evaluation of this model was conducted to ensure that the developed artifact addresses the problem identified in Chapter 1. The following sections describe the approach to ex-post evaluation that was used in this study.

Evaluation approach

Several approaches to design science evaluation focus on three quality the syntactic, semantic, and pragmatic (Helfert et al., 2012; Lei, 2012; Maes and Poels, 2007; Rittgen, 2010).

The first systematic framework to help evaluate quality in conceptual modelling was created by Lindland et al. in 1994. The framework takes quality into account not only in models but also in the conceptual modelling process. Figure 46 depicts the framework, and its components are detailed below.

- *Domain*: a collection of all correct and relevant statements that describe a problem domain.
- *Language:* the statements that a modelling language's syntax can make.
- *Model:* sets of statements that are made.
- Audience Interpretation: the collection of statements from a model that an audience believes it possesses.

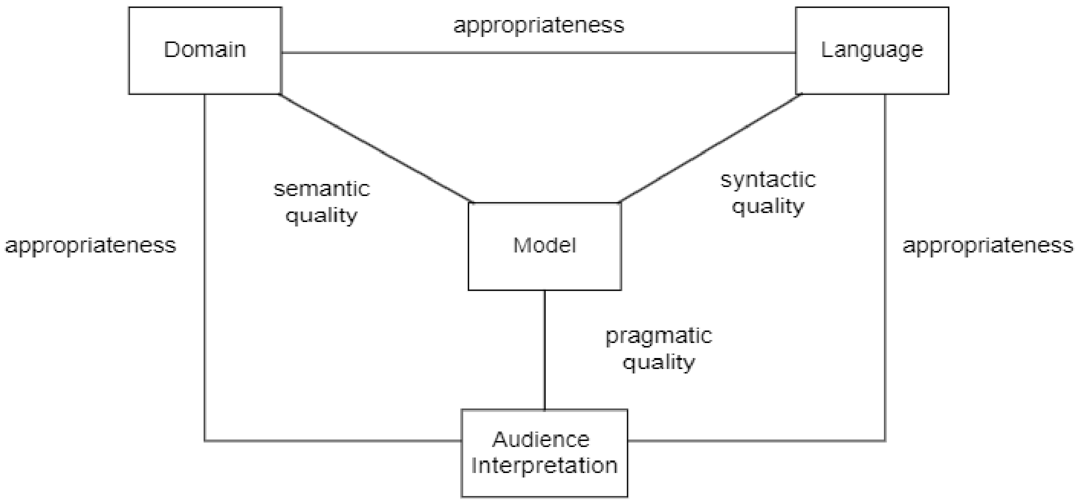

Figure 46 Lindland et al. (1994) Framework

This study uses the following three questions from Pries-Heje et al. (2008) strategic framework to guide its evaluation in this section (see Table 26):

1. What artifact is being evaluated?
2. How is the artifact is being evaluated (For instance, is it evaluated in a natural or artificial setting? What process and criteria are used to evaluate)?
3. When is the artifact evaluated? For instance, at what stage does the evaluation take place: ex-ante or ex-post)?

Table 26 Strategic Framework for Ex-post evaluation (Pries-Heje et al. 2008)

Questions	Relevance to the metamodel
What is being evaluated?	A data lifecycle metamodel
How is it evaluated?	A naturalistic evaluation is conducted to evaluate the perceived usefulness of the developed artifact
When is the evaluation taking place?	Ex-post, after the metamodel has been developed

5.3.3 Evaluation Criteria

Evaluation of a designed is a crucial activity in DSR. It is carried out to assure that the artefact serves its intended purpose (Hevner Chatterjee, 2010). Ex-post evaluation criteria for the syntactic, semantic, and pragmatic dimensions of an artifact are presented in Table 27 and the measurements statements thesis are presented in Table 28.

Syntactic Quality

The syntactic quality of a metamodel addresses the rules of a modelling language (Rittgen, 2010) and refers specifically to the metamodel's abstract syntax. The abstract syntax defines a model's concepts and attributes as well as their relationships (Bork and Fill, 2014). The criteria selected to evaluate the metamodel's syntactic quality are presented below.

Syntactic correctness. This criterion refers to the way statements in a model comply with its syntax (Rittgen, 2010). The metamodel developed in this study was specified in the meta object facility (MOF) standard during the design phase (see Section 4.2).

Semantic Quality

The semantic quality refers to concrete syntax and semantics. Concrete syntax defines a graphical notation that can represent the abstract syntax of a model (Cengarle et al., 2009). Semantics also refers to the interpretation of concepts and relationships presented in abstract syntax (Huber et al., 2019). The criteria selected to evaluate this dimension are presented below.

Easy to understand. This criterion is a graphical notation used to represent a metamodel (Huber et al., 2019) and to evaluate how well practitioners can understand the model's notation and semantics. In this study, the researcher used case studies to evaluate the model's understandability.

Completeness. This criterion refers to all statements that a metamodel makes about a domain that are correct and relevant (Rittgen, 2010). This means that all concepts used in the model possess the properties needed to describe and represent them (Bork and Fill, 2014). The metamodel's completeness in this study was also evaluated in the case studies.

Pragmatic Quality

The pragmatic quality is related to a user's interpretation of a model (Rittgen, 2010). It focuses on their comprehension of a model (Lei et al, 2012), and it assesses if a model fulfils a user's needs or its utility (Lei et al, 2012). The criterion selected to evaluate this dimension is presented below.

Relevance. This criterion assesses if all elements of a design are relevant to the problem (Rittgen, 2010). In this study, the metamodel was evaluated to ensure that it is relevant to smart city practitioners. Its relevance was also evaluated by the case studies.

Dimension	What to Evaluate	Criteria	Methods	Supporting Source
Syntactic	Abstract Syntax ▪ Metamodel concepts ▪ Relationships between concepts	Syntactical Correctness	Specify metamodel based on MOF	Bork and Fill (2014); Cengarle et al. (2009); Helfert et al. (2012); Huber et al. (2019); Rittgen (2010)
Semantic	Meaning of each syntactic element	▪ Easy to understand ▪ Completeness	Case Study (Cork and Limerick cities)	Bork and Fill (2014); Cengarle et al. (2009); Helfert et al. (2012); Huber et al. (2019); Rittgen (2010)
Pragmatic	Usefulness (utility) of the metamodel	Relevance	Case Study (Cork and Limerick cities)	Bork and Fill (2014); Helfert et al. (2012); Rittgen (2010)

Table 27 Ex post Evaluation Criteria

Table 28 Measurement statement in this thesis

Quality Criteria	Description	Statement to be measured
Easy to understand	The model is understandable for managers.	The graphical notations are easy to understand by smart cities practitioners.
Completeness	The representation contains all statements about the domain that are correct and relevant.	Defined attributes represent the necessary properties to describe the proposed concepts.
Relevance	All statements in the representation are relevant to the problem.	All concepts and relationships in the metamodel are relevant to represent the variations in data lifecycles in order to support smart city strategies.

5.3.4 Ex-Post Evaluation Results

The main purpose of the ex-post evaluation is to demonstrate the relevance and utility of the concepts that were identified and used to create the metamodel. This section examines the evaluation results. Meetings and semi-structured interviews with Limerick and Cork City Council practitioners were held to examine the proposed artifact. For the evaluation, four people from Irish city councils were interviewed. Their knowledge ranges from senior managers to managers with numerous years of public service experience. The researcher also shared a long iteration of the design process with the Cork City Council, presenting developments in the research to one of the managers on multiple occasions. This offered the chance to get their feedback and better understand the city council's issues and concerns.

Table 29 contains a transcription of an interview with a manager from the Cork City Council, who provided feedback on the proposed concepts and their use.

Table 29 Detailed Expert Feedback

Concept	Transcription from Interview	Action Taken
Objective	A: "The objective is important to us." The city provides many services and processes a large quantity of data, but we need to process data in different ways, so defining objectives can help us determine whether we are meeting our goals."	A: "We can confirm the relevance of the definition of the concept objective."
Phases / Activities	A: "The phases and activities of a data lifecycle do not occur in a linear way. We have structured and unstructured data and data evolves as the process progresses. So, it is vital to plan how data is acquired, integrated, secured. As the systems/processes are audited by a data processor office, data managers need to know how data is processed. And we must adhere to the data strategy's guiding principles such as transparency, accessibility and reusability."	A: "We can confirm the relevance of the definition of these concepts."
Variation Driver	A: "This is a very important concept for us. We use several types of data lifecycles because we need to process data in various ways to suit various requirements. We must simultaneously adhere to regulations and be knowledgeable about diverse requirements. The model clearly demonstrates the concepts that influence phase and activity decisions. It appeals to me since it allows us to include more concepts as part of the variation driver. We need some flexibility in a model to meet our demands because things are constantly changing and we must be prepared to adapt."	A: "We can confirm the relevance of the definition of this concept and the need of city managers to know how the variation driver impacts on the phases and activities of a data lifecycle."

Concept	Transcription from Interview	Action Taken
Data Type	A: "As previously stated, we have various types of data, and we must consider this information in order to know how to process it and what steps are necessary to accomplish our goals."	A: "We can confirm the relevance of the definition of this as it is used to make decisions."
Processing Type	A: "We need to process data in many ways, and the ability to process data quickly is essential in some services. While some services must be processed in real-time, others don't. One of our aims is to increase connected services for citizens so we need an effective system to use data in an efficient way."	A: "We can confirm the relevance of the definition of this as it is used to make decisions."
Processing Location	A: "Our City Council does not process all data. Some of them are handled by third parties with whom we have a data processing agreement. With this concept we can have more information in a model."	A: "We can confirm the relevance of the definition of this as it is used to make decisions."
Data Lifespan	A: "I think this concept is very important, because we don't delete data because we are afraid we will need it in the future, but we must pay attention to this. I believe we need to change our approach on this matter, change our mindset."	A: "We can confirm the relevance of the definition of this concept."

Table 29 Continuation

Concept	Transcription from Interview	Action Taken
Category / Sensitivity	A: “These concepts are among the most crucial to us. Data processing must be done in a secure and confidential manner. Unauthorised individuals or entities cannot access data. Due to the problem Limerick City Council was experiencing, we had to review all of our procedures and to make sure that data was being collected, and used for the intended purpose and that it was accessible to the to appropriate parties.”	A: “We can confirm the relevance of the definition of these concepts.”
Regulations	A: "We are aware that in order to process data, we must adhere to regulations and exercise extreme caution. So, this concept cannot be left out of the model because it has an impact on how we have to handle data. Since the implementation of GDPR, we must be cautious about how we process data, as well as how third parties process data. Regulations are critical in data processing."	A: “We can confirm the relevance of the definition of this concept.”
Data Controller	A: "To us, a data controller is crucial. Since we handle personal data, the model must include this concept. The data controller must guarantee that data is processed in accordance with GDPR requirements. And, in accordance with our data strategy, we must process data in a way that is transparent to citizens.	A: “We can confirm the relevance of the definition of this concept.”
Pre/Pos Requirement	A: “The public service data strategy proposes the "Once-only" approach, which encourages data reuse by requiring citizens and organisations to contribute data only once and sharing and reusing data only when necessary. I believe these concepts are relevant; for example, the pre requirement might be used to determine whether a given data set has already been acquired before attempting to collect it again."	A: “We can confirm the relevance of the definition of these attributes.”

Syntactic Quality

Syntactic correctness. The syntactic correctness of the metamodel was evaluated using the ArchiMate enterprise modelling language. The modelling tool contains a model validator that verifies if a model complies with ArchiMate specifications. The model has passed the test.

Semantic

Easy to understand. The graphical notation of the metamodel's concepts and their visual representation were analysed by practitioners during the interviews. Each scenario was analysed, and user feedback was positive. The notation used to represent the proposed concepts was easy to understand.

Completeness. The practitioners stated that the metamodel contains all statements relating to the smart city domain and that they are correct and relevant. They confirmed that the properties and attributes used to represent concepts are necessary.

Pragmatic Quality

Relevance. Practitioners highlighted the relevance of the problem addressed by this research. They stated that the concepts and their relationships will assist them in modelling variations in data lifecycles in smart city architectures. They acknowledged the importance of models in helping them communicate with other stakeholders and in understanding various views and aspects of the domain.

Table 30 details practitioner comments on the relevance of the metamodel. The overall response from practitioners was positive, and the model was validated. Below are comments that practitioners made about the quality criteria used to evaluate the metamodel.

Table 30 Detailed expert feedback- Relevance

Topic	**Transcription from Interview**
Concepts and Relationships:	A: The concepts suggested are useful for us to use on a daily basis and they can assist to have a common language to discuss with other stakeholders. B: The proposed concepts are very relevant to represent a complicated field as smart cities and it is a good contribution to support our work in the city council. It also gives us flexibility to add elements to the variation driver accordingly with our needs. For instance, the connection between data lifespan, and regulation to the variation driver reflects a connection with public data strategy that we need to follow.
City Service example:	The example that you used was easy to understand and very relevant to this city council, as other city council had a serious problem regarding this service.

5.3.5 Measures for this Research

Key Performance Indicators (KPIs) are quantifiable measures that indicate how well organizations are performing in meeting their objectives. Organizations use them to monitor, assess, and manage their performance. KPIs are classified into numerous categories, including sales, customer service, information technology, finance, and human resources (van de Ven et al., 2023). KPIs were utilised to assist in evaluating the results of this research, and they are shown below.

KPI: Completeness of Requirements

This KPI determines if the proposed metamodel captures the fundamental concepts required by practitioners. Failure to meet conditions, as indicated by Dragicevic et al. (2014), may result in an ineffective artifact.

KPI: Satisfaction of customer needs

This KPI measures the extent to which the proposed metamodel meets the requirements or needs of the practitioners and it has a qualitative scale from high to low (van de Ven et al., 2023).

Table 31 presents the results of the KPIs collected on interviews conducted with practitioners from Cork and Limerick city councils.

Table 31 KPIs for this thesis

KPI	Description	Result
Completeness of Requirements	Assessing if the proposed metamodel captures the concepts necessary to model data lifecycles in the smart city domain.	95%
Satisfaction of customer needs	Assessing if the proposed metamodel meets the needs of practitioners to model the services provided by the city councils.	High

5.3.6 Validity and Reliability of this Research

This study has applied four tests to assess the quality of its research design: construct validity, internal validity, external validity, and reliability.

Construct validity refers to the correct identification of a set of measures regarding concepts that are going to be studied (Yin, 2014). Data triangulation was used in this study to increase construct validity, as triangulation increases the precision of empirical research (Runeson and Höst, 2009). To achieve triangulation, several data sources were used, including meetings, semi-structured interviews, city council websites, and reports. According to Yin (2009), construct validity can also be increased by creating a case study report for different audiences. Thus, a conference article and a book chapter of a case study were accepted, and the main research findings were validated by peer reviewers.

Internal validity refers to the investigation of cause and effect between events (Yin, 2009). To address this aspect of the metamodel, the researcher maintained a long-term collaboration with an ISO standard group and a long-term research relationship with a manager from the Cork City Council. These collaborations allowed the researcher to understand data classification issues faced by organisations, as well as city services and data management issues faced by city councils in Ireland. The researcher also followed enterprise architecture guidelines and modelling techniques to develop the metamodel. The concepts and their relationships were extracted from literature and practice.

External validity refers to how research results can be generalised apart from the research methods used (Yin, 2014). According to Runeson and Höst (2009), conducting multiple case studies assists in the generalisation of findings. To guarantee the metamodel's external validity, the researcher conducted case studies in two cities in Ireland. The objective was to confirm the relevance of the research findings. During meetings and interviews, practitioners emphasised the model's relevance for helping them model variations in their data processing.

Reliability ensures the repeatability of a research outcome (Rittgen, 2010; Yin, 2014). In this study, the metamodel's reliability was guaranteed by using the modelling method, by following standard guidelines, and by creating a case study protocol. From the beginning, the research involved clear methods to ensure its reliability. Data triangulation was used in both case studies to validate findings. Protocols were also observed when preparing the interviews and during data analysis. The development of the metamodel followed the MDE approach and AMME as presented in Section 4.2.

5.4 Summary

This chapter presented the demonstration and evaluation of the proposed metamodel. The demonstration showed how the metamodel functioned in two case studies. The evaluation assessed how useful the metamodel was for practitioners from Cork and Limerick. The evaluation was split into ex-ante and ex-post phases, and the quality of the model was assessed on three levels: syntactic, semantic, and pragmatic. The evaluation results confirmed the relevance of the research problem and the model for smart city practitioners.

The next and final chapter provides a review of this study, including its contribution, research limitations, and future work.

Chapter 6

6. Conclusion

This research investigated the modelling of data lifecycles in enterprise architectures. It identified the limitations of current modelling of this tool in the smart city context and proposed a metamodel to model data lifecycle variations in the smart city domain. The metamodel defines syntax and semantics, allowing smart city stakeholders to communicate in a common language. This study evaluated the metamodel using ex-ante and ex-post approaches.

This chapter begins by summarising the study's research questions and findings. It then discusses the contribution that the thesis may make to both academia and practice. The chapter ends by discussing the study's limitations and how the study may be improved in the future.

6.1 Revisiting Research Questions

Data plays a significant role in helping business and organizations make better decisions and provide better services and products to customers. Managing data, however, can be very challenging. This research addresses the challenge of modelling data lifecycle variations within enterprise architectures, and it demonstrates its findings in the smart city context.

The research questions identified to address these challenges are presented in Section 1.4. This section revisits the RQs and presents their main results.

In order to design a solution that addresses the problem identified by this research, the main research question – *"How to support the representation of variations in data lifecycles in enterprise architectures in the context of smart cities?"* – was divided into three sub-research questions.

For Research Question1: To answer this research question, a literature review was conducted and presented in Chapter 2. The result was the identification of elements necessary to model a data lifecycle.

- Eight elements were found and identified as main elements for modelling a data lifecycle.
- The literature review identified four additional elements that must be considered, as they influence how data is handled. These elements are data lifespan, category, sensitivity, and regulations; each directly impacts the choice of phases and activities in a data lifecycle.

For Research Question 2: Chapter 4 outlines how a metamodel, using the elements identified in RQ1, was developed to model variations in data lifecycles within enterprise architectures.

- Four design requirements were identified for defining the functionalities that the metamodel must have. These design requirements were identified in the literature review and validated with smart city domain experts.
- Variation options in data lifecycles were identified that would be represented by a variation driver.
- Smart city concepts and their relationships were defined and represented in the metamodel. The specification of the metamodel includes its syntax and semantics.

For Research Question 3: Chapter 5 outlines how the metamodel can be used and defines an approach for evaluating its concepts and relationships.

- The relevance and applicability of the identified concepts and their relationships were demonstrated through case studies in two cities in Ireland.
- The quality of this research was assessed using validity and reliability tests.

6.2 Thesis Contribution

The literature review in Chapter 2 discusses the value of using data lifecycle models for organisations, the variety of models available, and the drawbacks of these models.

The main result of this research is a metamodel that identifies new concepts and their relationships for modelling variations in data lifecycles and for assembling a coherent enterprise architecture. The new concepts allow city managers to:

- Manage a data lifecycle: objective, phases/activities

- Manage variability in data lifecycles: variation driver
- Manage data requirements: data type, processing type, processing location and data lifespan.
- Classify data: category, sensitivity and policies/regulations.

This study addresses gaps in data lifecycle modelling by investigating the variations necessary in data lifecycles and improving their modelling. The researcher identified concepts to model data lifecycle variations in enterprise architecture in the smart city context (section 4.2.3). Research also addressed why these variations occur and what to consider when they occur (section 2.2.3).

The researcher decided to conduct this study in the smart city context because of its complexity: it offers an ideal scenario in which different data lifecycles must be used depending both on how data needs to be processed and on privacy and security requirements. Another result of this research is the development of a data taxonomy (section 4.4). The taxonomy was developed to understand data specifications from the smart city domain and to assist in identifying new concepts for modelling data lifecycle variations.

In summary, the proposed metamodel offers practitioners a common language to assist in their communication. The next sections will outline the impact that this thesis contribution may have for researchers and practitioners.

6.2.1 Theoretical Contribution

This study's main contribution is a novel artifact as well as the introduction of new concepts. According to Gregor and Hevner

(2013)'s DSR Knowledge Contribution Framework, the contribution of this thesis is classified as Exaptation as indicated in Figure 47. This type of contribution extends or improves design knowledge in one field to be used in a new application area. The researcher must demonstrate that the new extension of design knowledge is interesting and some specific challenges are presented in the new field.

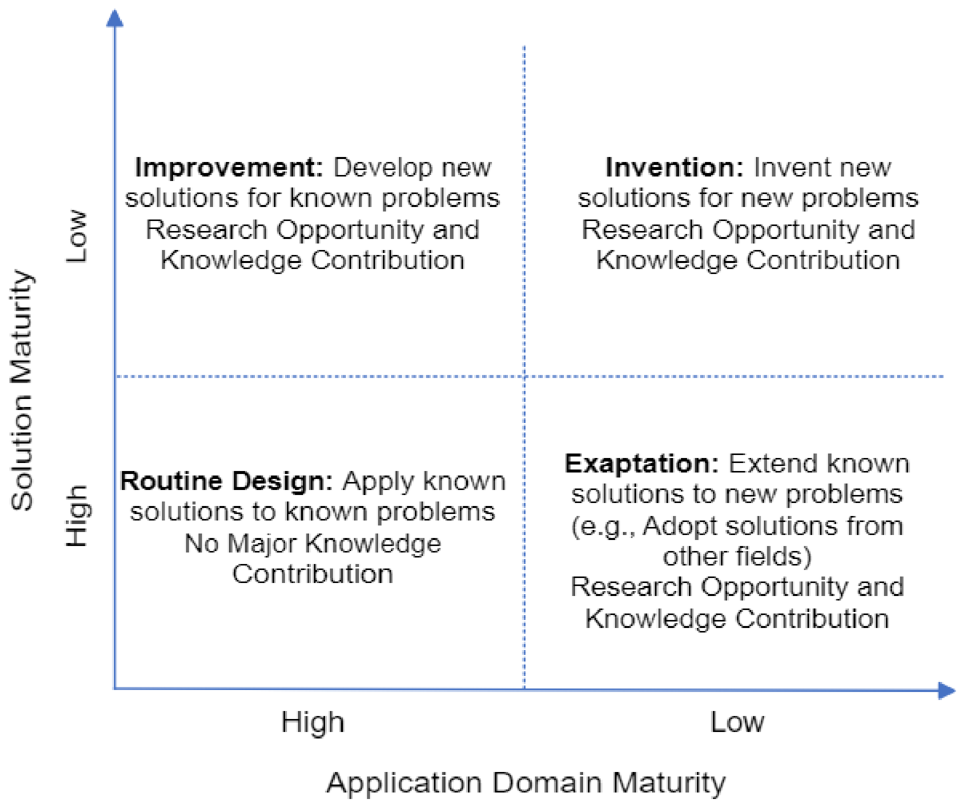

Figure 47 DSR Knowledge Contribution Framework (Gregor and Hevner, 2013)

Existing research focuses mostly on the data analysis phase, with limited emphasis on data lifecycle modelling. The importance of using data lifecycles for data processing in smart cities was demonstrated in Section 2.1. This research focuses on modelling data lifecycles and ways to improve the modelling of this data management tool. This study lists the main elements of this tool that were identified in the literature review in section 2.2.3. One of the contributions of this research is the addition of important elements that must be addressed when processing data, including data lifespan, regulations, category, and sensitivity. These elements impact the choice of phases and activities in a data lifecycle. A smart city offers different types of services to citizens and different data processing requirements are required to process data.

As stated in Section 2.5, EA metamodels can help practitioners to handle complexity in a smart city context however, the various existing models lack the necessary concepts to model data processing requirements. This research uses the concept of a variation driver to represent these requirements, which enables the representation of variations that occur in data lifecycles. This research's main result is a metamodel, which is a contribution to the field of enterprise architecture modelling.

This metamodel contains concepts that enable the modelling of required variations in data lifecycles, allowing smart city practitioners to model varied data processing based on their needs and requirements.

Another contribution of the research is an understanding of data. Section 4.2.2 depicts a taxonomy created to comprehend data specifications from the smart city domain. This study revealed aspects that must be considered when processing data, which can vary depending on the requirements. flows in the smart city domain.

To summarize, this study's contributions enable the modelling of data lifecycle variations in EAs to guarantee the representation of various data processing requirements in the context of smart cities.

6.2.2 Impact for Research

Recent research has focused mostly on data analysis and the technical challenges faced by smart cities, but it has also helped us understand the challenge of modelling data lifecycles within EA in the context of smart cities.

The current study contributes to research by improving our understanding of how data modelling and processing should align with city data strategies (see Section 2.1).

This study also helps define the concepts and relationships needed to model a data lifecycle within EAs in the smart city domain. One such concept that was identified in the study is the variation driver, which directly impacts the definition of phases and activities needed to process data and provide services to citizens. The variation driver helps model a data lifecycle by accounting for processing requirements such as data classification. Concepts like the variation driver contribute to the development of a coherent architecture that helps city planners and data managers handle the complexity of data management in smart cities.

In the two case studies conducted in this study, the researcher developed models that illustrate how different cities require different configurations of the proposed concepts and relationships that accord with their specific data processing requirements and city strategies.

The main contributions for research are listed below.

For Research Question 1: this researcher combined the findings of literature review with the perspectives of practitioners to identify elements of data lifecycles.

For Research Question 2: It developed a metamodel to model variations in data lifecycles.

- It developed a data taxonomy to comprehend data specification from the smart city domain.
- Identified four design requirements to define functionalities that the metamodel should have in order to achieve the research objectives (see Section 1.5).

For Research Question 3: It examined the developed artifact both during and after development (ex-ante and ex-post). Additionally, demonstrated how the proposed metamodel might be applied to real-world case studies.

Overall, the findings of this study enrich academic discussion by improving our understanding of how the identified concepts are linked and work together. The metamodel may be used as a reference to guide the creation of a coherent EA and to compare various data lifecycles needed to meet different data processing requirements.

6.2.3 Impact for Practice

The research in this study also benefits practice. By examining numerous data lifecycles, the researcher was able to understand their differences, the reasons for their differences, and how to compare different data management tools. This knowledge was then used to develop the proposed metamodel, which includes concepts and relationships that are not present in previous models but are regarded as important to practitioners.

Practitioners can model various data lifecycles within EA using the proposed metamodel as a guide. It can provide them with a holistic view of data processing for city services, enabling them to identify areas for improvement and ensure that data processing accords with regulations and other requirements. The introduction of the variation driver can help practitioners understand different data lifecycles and also compare them through variations points and variations options (see Table 24).

The two case studies validated the relevance of this research to city councils (see Section 4.2.1). If practitioners are to provide services in a smart city, they need to plan how the city's data will be processed. This includes processing data in various ways and adhering to various processing requirements, such as complying with regulations; it also includes providing transparency to citizens by informing them how their data is being processed and used. Until now, practitioners did not have a model at their disposal that contained all of this information to aid their decision-making.

The research also showed that the proposed metamodel may be adapted to model different requirements, such as data strategy and regulations, that must be considered while processing data in a smart city.

Overall, the research in this study helps us understand the variations of data lifecycles, why they occur, and what factors to consider while modelling them.

Data lifecycles might vary depending on the needs of data management. These variations occur for a variety of reasons including the diversity of data sources, diversity of processing requirements and data strategies. Effective data management is essential not just for maximising data value, but also for ensuring data security and privacy.

Therefore, these models in turn can help stakeholders address changing requirements and data management in smart cities.

6.3 Research Limitations

This research focuses on modelling data lifecycle variations with EAs. Due to time constraints and resources, it has some limitations. This section discusses these research limitations from a critical perspective.

Number of services

Smart cities offer services to their citizens in order to improve their lives. Data plays a strategic role in delivering these services and improving city decisions (Kim, 2022; Lau, 2019; Sarker, 2022). Number of services in a city can expand and a good data management is necessary to manage them. This study identified from the literature how data lifecycles vary and case studies proved the variation driver's applicability to practitioners. The case studies were conducted in cities with a small number of services, but larger cities may provide a wide range of services to their citizens. Therefore, it is important to evaluate the proposed metamodel in large-scale cities and determine whether their services require concepts not uncovered by this research.

Cross domain solutions

A smart city is divided into domains: health care, mobility and transportation, and environment, to name a few (Lau et al., 2019; Sánchez-Corcuera et al., 2019). Each domain may offer a variety of services to citizens. Literature shows that exchanging information about services between different domains can improve interoperability among the city's systems and benefit multiple stakeholders (Hefnamy et al., 2015). Due to time constraints, this study did not have the opportunity to explore solutions for cross-domain services. In the Cork case study, research focussed on the CCTV service in the security domain, since the city had some problems with this service. In the Limerick case study, research focussed on the footfall counter service in the movement and transportation domain. Neither situation offered an opportunity to explore implemented solutions in multiple domains. Future research is therefore needed to design and evaluate the metamodel for cross-domain solutions.

Covid-19

The emergence of COVID-19 required the researcher to adjust to new methods for conducting research. In particular, interviews had to be conducted online rather than on-site, which would have provided better interaction and evaluation. COVID also postponed the start of research evaluation, although the researcher was still able to meet with practitioners from several city councils to conduct meetings and interviews.

6.4 Future Work

There are a few options available for future work. As mentioned in the previous section, future research is needed to validate the proposed metamodel in cities with a higher number of services, and it is needed to design cross-domain solutions to demonstrate and evaluate the flow of information between city domains. Such research will provide an opportunity to enhance the metamodel by adding more concepts, such as data consent and data provenance. The proposed metamodel was used in the context of a smart city to address its data management challenges. However, its application is not restricted to the smart city domain, as several other domains have challenges related to data management. The metamodel may therefore be used in other domains to analyse its effectiveness and assistance to stakeholders and to verify the need to add new concepts.

The inclusion of concepts such as data consent and data provenance is crucial to improve the modelling of data lifecycles. Since data provenance is seen as a component of a data lifecycle, data lifecycles and data provenance are related concepts. Data provenance information is gathered at various stages of a data lifecycle therefore integrating these two concepts improves data traceability and accountability, allowing for the identification of error sources and availability of reliable data. Data consent has a significant impact on phases of a data lifecycle such as collection, security, storage, data processing, and deletion, when particular types of data, such as personal data, must be collected.

The study's research findings were validated in two case studies, in which practitioners stated that the metamodel's concepts and conceptual relationships had practical relevance for their cities. As a result, another proposal is to extend ArchiMate in order to support the modelling of data lifecycle variations. ArchiMate is an open modelling language for EA, and anyone can contribute to the language evolution and the development of its Specification. The implementation of the identified concepts in ArchiMate would likely facilitate the representation of EA and assist practitioners in making decisions but also will help them to validate their models.

Other extensions of this research include exploring further the variation options, which can be optional, mandatory, or alternative. The proposed metamodel may also be expanded to add constraints. Constraints can be related to business (e.g., cost or budget constraints) or to IT (e.g., the constraint posed by integration with a software system). In this study, sensitive information was identified that needed to be deleted using specific procedures to prevent it from being recovered. It would be worthwhile to develop the research in this direction, as sensitive information may impose requirements that can constrain data management options.

Regarding publications, I am currently working on an article with other colleagues in which we propose a data fabric model to comprehend the relationship between data lifecycle, data value, data provenance, and consent management under an overarching concept of data governance. The plan is to submit this article to the 32nd European Conference on Information Systems in June 2024.

There is also the goal of writing another article to present the metamodel I developed in this study and address the significance of modelling data lifecycles and their variations. This will enable data lifecycles to be dynamically modelled to accommodate the evolving requirements that cities must consider to provide services to their citizens. The article will be submitted to the Journal Online Information Review in May 2024.

References

Abu Matar, M. and Mizouni, R. (2018). Variability Modeling for Smart City Reference Architectures. 10.1109/ISC2.2018.8656967.

Ahmad, K., Maabreh, M., Ghaly, M., Khan, K., Qadir,J. and Al-Fuqaha, A., (2022) Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges, Computer Science Review, Volume 43, 100452, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2021.100452.

Ahvenniemi, H., Huovila,A., Pinto-Seppä, I., and Airaksinen,M. (2017). What are the differences between sustainable and smart cities?, Cities, Volume 60, Part A, Pages 234-245, ISSN 0264-2751, https://doi.org/10.1016/j.cities.2016.09.009. (https://www.sciencedirect.com/science/article/pii/S0264275116302578).

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

Alshammari, M. and Simpson, A. (2018). Personal Data Management: An Abstract Personal Data Lifecycle Model. 10.1007/978-3-319-74030-0_55.

Anthopoulos, L. and Tsoukalas, I. (2006). The implementation model of a digital city. The case study of the digital city of trikala, greece: e-trikala. Journal of e-Government,2(2):91–109.

Anthopoulos, L. and Fitsilis, P. (2014). Exploring architectural and organizational features in smart cities. In International Conference on Advanced Communication Technology (ICACT), 2014 16th, pages 190–195. IEEE.

Arksey, H. and O'Malley, L. (2005) Scoping studies: towards a methodological framework, International Journal of Social Research Methodology, 8:1, 19-32, DOI: 10.1080/1364557032000119616

Atkinson, Colin; Kühne, Thomas (2020): A Deep Perspective on the ArchiMate Modeling Language and Standard. Enterprise Modelling and Information Systems Architectures (EMISAJ) – International Journal of Conceptual Modeling: Vol. 15, Nr. 2. DOI: 10.18417/emisa.15.2. Berlin: Gesellschaft für Informatik e.V.. pp. 1-25

Attaran, H., Kheibari, N., and Bahrepour, D. (2022). Toward integrated smart city: a new model for implementation and design challenges. *GeoJournal*, *87* (Suppl 4), 1–16. Advance online publication. https://doi.org/10.1007/s10708-021-10560-w

Attard, J., Orlandi, F. and Auer, S. (2016) 'Data driven governments: Creating value through open government data', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9860 LNCS(September), pp. 84–110. doi: 10.1007/978-3-662-53416-8_6.

Bankauskaite, J. (2019). Comparative analysis of enterprise architecture frameworks. *International Conference on Information Technology.*

Barat, S., Kulkarni, V., Clark, T., Barn, B. (2016). Enterprise Modeling as a Decision Making Aid: A Systematic Mapping Study. In: Horkoff, J., Jeusfeld, M., Persson, A. (eds) The Practice of Enterprise Modeling. PoEM 2016. Lecture Notes in Business Information Processing, vol 267. Springer, Cham. https://doi.org/10.1007/978-3-319-48393-1_2

Barnaghi, P., Bermúdez-Edo, M. and Tönjes, R. (2015). Challenges for Quality of Data in Smart Cities. Journal of Data and Information Quality. 6. 1-4. 10.1145/2747881.

Barrutia, J.M., Echebarria, C., Aguado-Moralejo, I., Apaolaza-Ibáñez, V. and Hartmann, P. (2022) Leading smart city projects: Government dynamic capabilities and public value creation, Technological Forecasting and Social Change, Volume 179, 121679, ISSN 0040-1625,

https://doi.org/10.1016/j.techfore.2022.121679.

Basias, N., and Pollalis, Y. (2018). Quantitative and qualitative research in business & technology: Justifying a suitable research methodology. *Review of Integrative Business and Economics Research, 7*, 91-105. Retrieved from https://may.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/quantitative-qualitative-research-business-amp/docview/1969776018/se-2

Bastidas, V., Bezbradica, M., Helfert, M. (2017). Cities as Enterprises: A Comparison of Smart City Frameworks Based on Enterprise Architecture Requirements. In: Alba, E.

Bastidas, V., Reychav, I., Ofir, A., Bezbradica, M. and Helfert, M. (2022) Concepts for Modeling Smart Cities. *Bus Inf Syst Eng* 64, 359–373. https://doi.org/10.1007/s12599-021-00724-w

Bézivin, J. (2004). In search of a basic principle for model driven engineering. Novatica Journal, Special Issue, 5(2):21–24.

Bifulco, F., Tregua, M., Amitrano, C., D'Auria, A. (2016). ICT and sustainability in smart cities management. Int. J. Public Sector Manag. 29, 132–147

Bohli, J., Skarmeta. A., Moreno, V., García,D. and Langendörfer,P. (2015) SMARTIE project: Secure IoT data management for smart cities, *International Conference on Recent Advances in Internet of Things (RIoT)*, Singapore, 2015, pp. 1-6. doi: 10.1109/RIOT.2015.7104906

Bork, D. (2018). Metamodel-Based Analysis of Domain-Specific Conceptual Modeling Methods. 172-187. 10.1007/978-3-030-02302-7_11.

Bork, D., Karagiannis, D. and Pittl, B. (2019). A survey of modeling language specification techniques. Information Systems. 87. 101425. 10.1016/j.is.2019.101425.

Bossaller, J. and Million, A. J. (2023). The research data life cycle, legacy data, and dilemmas in research data management. *Journal of the Association for Information Science and Technology*, 74 (6), 701– 706. https://doi.org/10.1002/asi.24645

British Standard (2017) BS 10010:2017, Information classification, marking and handling.

Cabinet Office. Government Security Classifications. (2018a).

Cabinet Office. International Classified Exchanges. (2018b).

Carlson, J., (2014) The Use of Life Cycle Models in Developing and Supporting Data Services. In J. M. Ray (Ed.), *Research Data Management. Practical Strategies for Information Professionals.* West Lafayette: Purdue University Press., 63-86.

Carnegie Mellon University. (n.d.a) Information Security Office. [Available at https://www.cmu.edu/iso/governance/guidelines/data-classification.html]

Castelli, M., Comerio, M. and Cremaschi, M. (2013). Towards the Definition of Value-added Services for Citizens: a New Model for the Description of Public Administration Services. International Journal of Management & Information Technology. 4. 166-173. 10.24297/ijmit.v4i1.4635.

Chauhan, S., Agarwal, N. and Kar, A.K. (2016). Addressing big data challenges in smart cities: a systematic literature review. *info.*

Chen, C.-W. (2023). Can smart cities bring happiness to promote sustainable development? Contexts and clues of subjective well-being and urban livability, Developments in the Built Environment, Volume 13, 100108, ISSN 2666-1659, https://doi.org/10.1016/j.dibe.2022.100108.

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S. and Zhou, X. (2013). Big data challenge: A data management perspective. Frontiers of Computer Science. 7. 10.1007/s11704-013-3903-7.

Chen, M., Mao, S. and Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications. 19. 10.1007/s11036-013-0489-0.

Chicano, F., Luque, G. (2017) (eds) Smart Cities. Smart-CT. Lecture Notes in Computer Science(), vol 10268. Springer, Cham. https://doi.org/10.1007/978-3-319-59513-9_3

Chief Information Officer, U.S. Department of Defense (2010). Department of defense architecture framework version 2.02.

Chief Information Officers Council (1999). Federal Enterprise Architecture Framework Version 1.1.

Chiprianov, V., Kermarrec, Y., Rouvrais, S., and Simonin, J. (2014). Extending enterprise architecture modeling languages for domain specificity and collaboration: application to telecommunication service design. Software & Systems Modeling, 13(3):963–974.

Christopherson, L., Mandal, A., Scott, E. and Baldin, I. (2020). Toward a Data Lifecycle Model for NSF Large Facilities. 168-175. 10.1145/3311790.3396636.

Chun Tie Y, Birks M, Francis K. (2019) Grounded theory research: A design framework for novice researchers. SAGE Open Med. 2019 Jan 2;7:2050312118822927. doi: 10.1177/2050312118822927. PMID: 30637106; PMCID: PMC6318722.

Costa, C. and Santos, M. Y. (2016). Basis: A big data architecture for smart cities. In 2016 SAI Computing Conference (SAI), pages 1247–1256. IEEE.

Cork City website. Available at https://www.corkcity.ie/en/

Cox, A., and Tam, W., (2018) A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management, 70*(2), 142–157. https://doi.org/10.1108/AJIM-11-2017-0251.

Cox, A., Parslow, P., De Lathouwer, B., Klien, E., Kempen, B., and Lonien, J. (2016). Espresso project d4.2 – definition of smart city reference architecture, systemic standardisation approach to empower smart cities and communities. Last accessed 12 December 2019.

Creswell, J. W. (2015). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (5th ed.). Boston, MA: Pearson.

De Guimarães, J.C.F., Severo, E.A., Júnior, L.A.F., Da Costa, W.P.L.B., and Salmoria, F.T. (2020). Governance and quality of life in smart cities: Towards sustainable development goals, Journal of Cleaner Production, Volume 253, 119926, ISSN 0959-6526,

https://doi.org/10.1016/j.jclepro.2019.119926.

Demchenko, Y., Los, W. and Laat, C. (2018). DATA AS ECONOMIC GOODS: DEFINITIONS, PROPERTIES, CHALLENGES, ENABLING TECHNOLOGIES FOR FUTURE DATA MARKETS.

Demchenko, Y., de Laat, C., and Membrey, P. (2014). "Defining architecture components of the Big Data Ecosystem," *2014 International Conference on Collaboration Technologies and Systems (CTS)*, Minneapolis, MN, USA, pp. 104-112, doi: 10.1109/CTS.2014.6867550.

Department of Public Expenditure, NDP Delivery and Reform (2019). Public service strategy 2019-2023. https://www.gov.ie/en/publication/1d6bc7-public-service-data-strategy-2019-2023/

Digital Guardian. (n.d.a) The Definitive Guide to Data Classification.

Dragicevic, S., Celar, S. and Novak, L. (2014). "Use of Method for Elicitation, Documentation, and Validation of Software User Requirements (MEDoV) in Agile Software Development Projects," *2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks*, Tetova, Macedonia, pp. 65-70, doi: 10.1109/CICSyN.2014.27.

Dumitriu, D. and Popescu, M. (2020). Enterprise Architecture Framework Design in IT Management. Procedia Manufacturing. 46. 932-940. 10.1016/j.promfg.2020.05.011.

Eger, J. (2009). Smart growth, smart cities, and the crisis at the pump a worldwide phenomenon. I-WAYS J. E-Gov. Policy Regul. 32, 47–53

El Arass, M., Tikito, I. and Souissi, N. (2017). Data lifecycles analysis: Towards intelligent cycle, *Intelligent Systems and Computer Vision (ISCV)*, Fez, pp. 1-8.

El Arass, M., Tikito, I. and Souissi, N., (2018) An Audit Framework for Data Lifecycles in a Big Data context. 10.1109/MoWNet.2018.8428883.

Elmekki, H., Chiadmi, D. and Lamharhar, H., (2019) Open Government Data: Towards a comparison of Data Lifecycle models. In Proceedings of the ArabWIC 6th Annual International Conference Research Track (ArabWIC 2019). ACM, New York, NY, USA, Article 15, 6 pages. DOI: https://doi.org/10.1145/3333165.3333180.

EMC (2015). Data Science and Big Data Analytics. Discovering, Analyzing, Visualizing and Presenting Data. John Wiley & Sons.

Erl, T., Khattak, W. and Buhler, P. (2016). Big Data Fundamentals: Concepts, Drivers & Techniques. 1st Edition. Prentice Hall.

Eryurek, E., Gilad, U., Lakshmanan, V., Kibunguchy-Grant, A., and Ashdown, J. (2021). Data Governance: The Definitive Guide: People, Processes, and Tools to Operationalize Data Trustworthiness.

Eurocities. (2019) Data people cities, Eurocities citizen data, principles in action. Available at: http://nws.eurocities.eu/MediaShell/media/Data_people_cities_-_EUROCITIES_citizen_data_principles_in_action.pdf

Falconer, G. and Mitchell, S. (2012). Smart city framework. Cisco Internet Business Solutions Group (IBSG), pages 1–11.

Faundeen, J. and Hutchison, V. (2017). The Evolution, Approval and Implementation of the U.S. Geological Survey Science Data Lifecycle Model. Journal of eScience Librarianship. 6. e1117. 10.7191/jeslib.2017.1117.

Fernández Áñez, V., Fernández-Güell, J., and Giffinger, R. (2017). Smart City implementation and discourses: An integrated conceptual model. The case of Vienna. Cities. 78. 10.1016/j.cities.2017.12.004.

Fernández-López, M., Gomez-Perez, A., and Juristo, N. (1997). METHONTOLOGY: from ontological art towards ontological engineering. Engineering Workshop on Ontological Engineering (AAAI97).

Filipponi, L., Vitaletti, A., Landi, G., Memeo, V., Laura, G., and Pucci, P. (2010). Smart city: An event driven architecture for monitoring public spaces with heterogeneous sensors. In 2010 Fourth International Conference on Sensor Technologies and Applications, pages 281–286. IEEE.

Gao, Y., Janssen, M., & Zhang, C. (2023). Understanding the evolution of open government data research: towards open data sustainability and smartness. *International Review of Administrative Sciences*, *89*(1), 59–75. https://doi.org/10.1177/00208523211009955

García Díaz, V., Núñez Valdez, E., Espada, J., Pelayo García-Bustelo, B., Cueva Lovelle, J. and Marín, C. (2014). A brief introduction to model-driven engineering. 18. 127-142.

Gaur, A., Scotney, B., Parr, G., and McClean, S. (2015). Smart city architecture and its applications based on iot. Procedia computer science, 52:1089–1094.

Gessa, A. and Sancha, P. (2020). Environmental Open Data in Urban Platforms: An Approach to the Big Data Life Cycle, Journal of Urban Technology, 27:1, 27-45, DOI: 10.1080/10630732.2019.1656934

Gharaibeh, A., Salahuddin, M., Hussini, S., Khreishah, A., Khalil, I., Guizani, M. and Al-Fuqaha, A., (2017). Smart Cities: A Survey on Data Management, Security and Enabling Technologies. IEEE Communications Surveys & Tutorials. PP. 1-1. 10.1109/COMST.2017.2736886.

Gil-Garcia, J. R., Pardo, T. A., and Nam, T. (2015). What makes a city smart? Identifying core components and proposing an integrative and comprehensive conceptualization. Information Polity, 20(1):61–87.

Gregor, S. and Hevner, A. (2013). Positioning and Presenting Design Science Research for Maximum Impact. MIS Quarterly. 37. 337-356. 10.25300/MISQ/2013/37.2.01.

Hajjaji,Y., Boulila, W., Farah, I., Romdhani,I., and Hussain, A. (2021) Big data and IoT-based applications in smart environments: A systematic review, Computer Science Review, Volume 39, 100318, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2020.100318.

Hancock, D. R. and Algozzine, B. (2017). Doing case study research: A practical guide for beginning researchers. Teachers College Press.

Hefnawy, A., Bouras, A., and Cherifi, C. (2015). Lifecycle based modeling of smart city ecosystem. In The 14th International Conference on Information and Knowledge Engineering.

Hefnawy, A., Elhariri, T., Cherifi, C., Robert, J., Bouras, A., Kubler, S., and Främling, K. (2017). Combined use of lifecycle management and IoT in smart cities. 1-9. 10.1109/SKIMA.2017.8294112.

Hefnawy, A., Bouras, A., and Cherifi, C. (2016). Iot for smart city services: Lifecycle approach. In Proceedings of the International Conference on Internet of things and Cloud Computing, page 55. ACM.

Helfert, M., Bastidas, V., and Pourzolfaghar, Z. (2018). Digital and smart services - the application of enterprise architecture. In International Conference on Digital Transformation and Global Society, pages 277–288. Springer.

Helfert, M., Donnellan, B., and Ostrowski, L. (2012). The case for design science utility and quality-evaluation of design science artifact within the sustainable ict capability maturity framework. Systems, Signs and Actions: An International Journal on Information Technology, Action, Communication and Workpractices, 6(1):46–66.

Hevner, A. March, P., and Ram., (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625.

Hevner, A and Chatterjee, S. (2010). Design Research in Information Systems. Series: Integrated Series in Information Systems, Volume 22, Page 87 DOI: 10.1007/978-1-4419-5653-8_8

Hou, J., Qu, L., and Shi, W. (2019). A survey on internet of things security from data perspectives, Computer Networks, Volume 148, Pages 295-306, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2018.11.026.

Huber, R., Püschel, L., and Röglinger, M. (2019). Capturing smart service systems: Development of a domain-specific modelling language. *Information Systems Journal, 29*, 1207 - 1255.

ISO (n.d.a) ISO/IEC 27001:2005 - Information technology -- Security techniques -- Information security management systems -- Requirements. Retrieve from https://www.iso.org/standard/42103.html.

ISO (n.d.a) ISO/IEC 27005:2008 - Information technology -- Security techniques -- Information security risk management.

Jin, J., Gubbi, J., Marusic, S., and Palaniswami, M. (2014). An information framework for creating a smart city through internet of things. IEEE Internet of Things journal, 1(2):112–121.

Johnston, A. (2014). Rigour in research: theory in the research approach. European Business Review.

Karagiannis, D. (2015). Agile modeling method engineering. In Proceedings of the 19th Panhellenic Conference on Informatics, pages 5–10.

Karagiannis, D., Bork, D. and Utz, W. (2019). Metamodels as a Conceptual Structure: Some Semantical and Syntactical Operations. 10.1007/978-3-030-06234-7_8.

Karagiannis, D. *et al.* (2019). 'A metamodeling approach to support the engineering of modeling method requirements', *Proceedings of the IEEE International Conference on Requirements Engineering*, 2019-September, pp. 199–210. doi: 10.1109/RE.2019.00030.

Kim, J. (2022). Smart city trends: A focus on 5 countries and 15 companies, Cities, Volume 123, 103551, ISSN 0264-2751, https://doi.org/10.1016/j.cities.2021.103551.

Khaloufi, H., Abouelmehdi, K., Beni-hssane, A., and Saadi, M. (2018) Security model for Big Healthcare Data Lifecycle, Procedia Computer Science, Volume 141, Pages 294-301, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.10.199.

Korhonen, J. and Halén, M. (2017). "Enterprise Architecture for Digital Transformation," *2017 IEEE 19th Conference on Business Informatics (CBI)*, pp. 349-358, doi: 10.1109/CBI.2017.45.

Koshy, E., Koshy, V., and Waterman, H. (2011). *Action research in healthcare*. SAGE Publications Ltd, https://dx.doi.org/10.4135/9781446288696

Kothari, C. R. (2004). Research methodology: Methods and techniques. New Age International.

Krogstie, J., Lindland, O.I. and Sindre, G. (1995). 'Defining quality aspects for conceptual models' in Information System Concepts Springer, 216-231.

Kshetri, N. (2014). Big data′s impact on privacy, security and consumer welfare. Telecommunications Policy. 38. 10.1016/j.telpol.2014.10.002.

Kuk, G. and Janssen, M. (2011). The business models and information architectures of smart cities. Journal of Urban Technology, 18(2):39–52.

Lankhorst, M. (2017). Enterprise Architecture at Work. 10.1007/978-3-662-53933-0.

Lara, P., Sanchez, M., and Villalobos, J. (2019). Ot modeling: The enterprise beyond it. Business & Information Systems Engineering, 61(4):399–411.

Lau, B., Marakkalage, S., Zhou, Y., Hassan, N., Yuen, C., Zhang, M., and Tan, U. (2019). A survey of data fusion in smart city applications, Information Fusion, Volume 52, Pages 357-374, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.05.004

Lei, L. (2012). Current practice in measuring the quality of conceptual models: Challenges and research. International Journal of Plant Engineering and Management.

Lim, C., Kim, K., Kim, M., Heo, J., Kim, K., and Maglio, P.P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. *Int. J. Inf. Manag., 39*, 121-135.

Lim, C., Kim, K. and Maglio, P. (2018). Smart cities with big data: Reference models, challenges, and considerations. Cities. 82. 10.1016/j.cities.2018.04.011.

Lindland, O.I., Sindre,G. and Solvberg, A. (1994). "Understanding quality in conceptual modeling," in *IEEE Software*, vol. 11, no. 2, pp. 42-49, March, doi: 10.1109/52.268955.

Liu, X., Heller, A., and Nielsen, P. (2017). CITIESData: a smart city data management framework. Knowledge and Information Systems. 1-24. 10.1007/s10115-017-1051-3.

Lněnička, M. and Komárková, J., (2019). Developing a government enterprise architecture framework to support the requirements of big and open linked data with the use of cloud computing. International Journal of Information Management. 46. 124-141. 10.1016/j.ijinfomgt.2018.12.003.

Lněnička, M., Máchová, R., Komárková, J., and Čermáková, I. (2017). Components of Big Data Analytics for Strategic Management of Enterprise Architecture.

Lnenicka, M., Machova, R., Komarkova, J., and Pasler, M. (2017). Government enterprise architecture for big and open linked data analytics in a smart city ecosystem. In InternationalConference on Smart Education and Smart E-Learning, pages 475–485. Springer.

Maes, A. and Poels, G. (2007). Evaluating quality of conceptual modelling scripts based on user perceptions. Data & Knowledge Engineering, 63(3):701–724.

Meth, H., Mueller, B., and Maedche, A. (2015). Designing a requirement mining system. Journal of the Association for Information Systems, 16(9):2.

Mohamed, I., Jusoh, Y., Abdullah, R., Nor, R., Nor, H. and Affendey, L. (2019). MEASURING THE PERFORMANCE OF BIG DATA ANALYTICS PROCESS. Journal of Theoretical and Applied Information Technology.

Möller, K. (2013). Lifecycle models of data-centric systems and domains: The abstract data lifecycle model. Semantic Web. 4. 67-88. 10.3233/SW-2012-0060.

Moura, F and de Abreu e Silva, J (2019). Smart Cities: Definitions, Evolution of the Concept and Examples of Initiatives. In: Leal Filho W., Azul A., Brandli L., Özuyar P., Wall T. (Eds) Industry, Innovation and Infrastructure. Encyclopedia of the UN Sustainable Development Goals. Springer, Cham. ISBN: 9783319710594.

Moustaka, V., Vakali, A., and Anthopoulos, L. (2018). A Systematic Review for Smart City Data Analytics. ACM Computing Surveys. 51. 1-41. 10.1145/3239566.

Myers, M. and Newman, M. (2007). The Qualitative Interview in IS Research: Examining the Craft. Information and Organization. 17. 2-26.10.1016/j.infoandorg.2006.11.001.

Nam, T. and Pardo, T. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. In: Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times, pp. 282–291. ACM

Nickerson, R. C., Varshney, U. and Muntermann, J. (2013). 'A method for taxonomy development and its application in information systems', *European Journal of Information Systems*. Nature Publishing Group, 22(3), pp. 336–359. doi: 10.1057/ejis.2012.26.

NIST (2004) FIPS PUB 199, 'Standards for Security Categorization of Federal Information and Information Systems'. doi: 10.6028/NIST.FIPS.199.

OECD (2020) Measuring Smart Cities' Performance. Do smart cities benefit everyone?

Oktaria, D., Suhardi, S. and Budi K. N., (2017). Smart City Services: A systematic Literature Review. 10.1109/ICITSI.2017.8267944.

OMG (2016). OMG Meta Object Facility (MOF) Core Specification, Version 2.5.1.

Ostrowski, L. and Helfert, M. (2012). 'Reference model in design science research to gather and model information', *18th*

Americas Conference on Information Systems 2012, AMCIS 2012, 5, pp. 3490–3499.

Paskaleva, K., Evans, J., Martin, C., Linjordet, T., Yang, D and Karvonen, A. (2017). Data Governance in the Sustainable Smart City. Informatics. 4. 41. 10.3390/informatics4040041.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Plale, B. and Kouper, I., (2017). The Centrality of Data: Data Lifecycle and Data Pipelines. In M. A. Chowdhury, A. Apon and K. Dey (Eds.), Data analytics for intelligent transportation systems (pp. 91–111). Cambridge, MA: Elsevier Inc. Retrieved from https://doi.org/10.1016/B978-0-12-809715-1.00004-3.

Pohl, K., Böckle, G., and Linden, F.J.v.d.L (2005). *Software Product Line Engineering: Foundations, Principles and Techniques*, 1st ed. New York: Springer-Verlag Berlin Heidelberg.

Poltavtseva, M. (2019). Evolution of Data Management Systems and Their Security. 25-29. 10.1109/EnT.2019.00010.

Pouchard, L. (2015). Revisiting the Data Lifecycle with Big Data Curation. International Journal of Digital Curation. 10. 10.2218/ijdc.v10i2.342.

Pries-Heje, J., Baskerville, R. and Venable, J.R. (2008). 'Strategies for design science research evaluation'.

Ramaprasad, A., Sánchez-Ortiz, A., Syn, T. (2017). A Unified Definition of a Smart City. In: Janssen, M., *et al.* Electronic

Government. EGOV 2017. Lecture Notes in Computer Science(), vol 10428. Springer, Cham. https://doi.org/10.1007/978-3-319-64677-0_2

Recker, J., 2005. Conceptual model evaluation. Towards more paradigmatic rigor. *life*, *11*, p.17.

Rittel, H. W. J. and Webber, M. M. (1973) 'Dilemmas in a general theory of planning', *Policy Sciences*, 4(2), pp. 155–169. doi: 10.1007/BF01405730.

Rittgen, P. (2010). Quality and perceived usefulness of process models. In Proceedings ofthe 2010 ACM Symposium on Applied Computing, pages 65–72.

Rokis, K., Kirikova, M. (2023). An ArchiMate-Based Thematic Knowledge Graph for Low-Code Software Development Domain. In: Abelló, A., *et al.* New Trends in Database and Information Systems. ADBIS 2023. Communications in Computer and Information Science, vol 1850. Springer, Cham. https://doi.org/10.1007/978-3-031-42941-5_40

Rubin, H.J. and Rubin, I.S. (2005) Qualitative Interviewing: The Art of Hearing Data. 2nd Edition, Sage Publications, Thousand Oaks.

Rubinstein, I.S. (2013) Big Data: The End of Privacy or a New Beginning?, *International Data Privacy Law*, Volume 3, Issue 2, Pages 74–87, https://doi.org/10.1093/idpl/ips036

Runeson, P. and Höst, M. Guidelines for conducting and reporting case study research in software engineering. *Empir Software Eng* 14, 131–164 (2009). https://doi.org/10.1007/s10664-008-9102-8

Rurua, N., Eshuis, R. and Razavian, M. (2019) Representing Variability in Enterprise Architecture. *Bus Inf Syst Eng* 61, 215–227. https://doi.org/10.1007/s12599-017-0511-3

Sánchez-Corcuera, R., Nuñez-Marcos, A., Sesma-Solance, J., Bilbao-Jayo, A., Mulero, R., Zulaika, U., Azkune, G. and Almeida, A., (2019). Smart cities survey: Technologies, application domains and challenges for the cities of the future. International Journal of Distributed Sensor Networks, 15(6), p.1550147719853984.

Şanta, A-M I. 2022. "Prosumers—A New Mindset for Citizens in Smart Cities" *Smart Cities* 5, no. 4: 1409-1420. https://doi.org/10.3390/smartcities5040072

Santana, E. F. Z., Chaves, A. P., Gerosa, M. A., Kon, F., and Milojicic, D. S. (2017). Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture. ACM Computing Surveys (CSUR), 50(6):78.

Sarker, I. (2022). Smart City Data Science: Towards data-driven smart cities with open research issues, Internet of Things, Volume 19, 100528, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2022.100528.

Sebastian-Coleman, L. (2022). Meeting the Challenges of Data Quality Management (1st edition), Academic Press, ISBN: 978-0-12-821737-5

Shah, S.I.H., Peristeras, V. and Magnisalis, I. (2021) DaLiF: a data lifecycle framework for data-driven governments. *J Big Data* 8, 89. https://doi.org/10.1186/s40537-021-00481-3

Shahid, J., Ahmad, R., Kiani, AK., Ahmad, T., Saeed, S., and Almuhaideb, AM. (2022) Data Protection and Privacy of the Internet of Healthcare Things (IoHTs). *Applied Sciences*. 12(4):1927. https://doi.org/10.3390/app12041927

Shanks, G., Gloet, M., Someh, I., Frampton, K., and Tamm, T. (2018). Achieving benefits with enterprise architecture. The Journal of Strategic Information Systems. 27. 10.1016/j.jsis.2018.03.001.

Siddiqa, A., Hashem, I., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A. and Nasaruddin, F. (2016). A Survey of Big Data Management: Taxonomy and State-of-the-Art. Journal of Network and Computer Applications. 71. 10.1016/j.jnca.2016.04.008.

Silva, A. (2015). Model-driven engineering: A survey supported by A unified conceptual model. Computer Languages, Systems & Structures. 20. 10.1016/j.cl.2015.06.001.

Sinaeepourfard, A., Masip-Bruin, X., Garcia, J. and Marín-Tordera, E., (2015). A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges. *Technical Report (UPC-DAC-RR-2015–18).*

Sinaeepourfard, A., García, J., Masip-Bruin, X., Marín-Tordera, E., Yin, X., and Wang, C. (2016a). A data lifeCycle model for smart cities. *International Conference on Information and Communication Technology Convergence (ICTC)*, 400-405.

Sinaeepourfard,A., Garcia, J., Masip-Bruin,X. and Marín-Tordera, E. (2016b) A comprehensive scenario agnostic Data LifeCycle model for an efficient data complexity management, *IEEE 12th International Conference on e-Science (e-Science)*, Baltimore, MD, pp. 276-281. doi: 10.1109/eScience.2016.7870909

Sonnenberg, C. (2012) 'Design Science Research in Information Systems. Advances in Theory and Practice', 7286(August 2014). doi: 10.1007/978-3-642-29863-9.

Statista Research Department 2022, Statista, https://www.statista.com/statistics/871513/worldwide-data-created/, accessed 07 December. 2022

Stoyanovich, J., Howe, B. and Jagadish, H.V. (2020) 'Responsible Data Management', *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3474-3488. https://doi.org/10.14778/3415478.3415570

Sutherland, M. and Cook, M. (2017). Data-Driven Smart Cities: A Closer Look at Organizational, Technical and Data Complexities. 471-476. 10.1145/3085228.3085239.

Tene, O. and Polonetsky, J. (2012). Big Data for All: Privacy and User Control in the Age of Analytics. Northwestern Journal of Technology and Intellectual Property. 11.

The Open Group (2019). Archimate 3.0.1 specification.

The Open Group (2018). Togaf standard version 9.2.

Thomas, D. R., (2006) A General Inductive Approach for Analyzing Qualitative Evaluation Data. American Journal of Evaluation, 27(2), 237–246. https://doi.org/10.1177/1098214005283748.

Thuan, N., Drechsler, A., and Antunes, P. (2019). Construction of Design Science Research Questions. Communications of the Association for Information Systems, 44, pp-pp. https://doi.org/10.17705/1CAIS.04420

Urbaczewski, L., and Mrdalj, S. (2006). A comparison of enterprise architecture frameworks. *Issues in Information Systems, 7*(2), 18–23.

Urbinati, A., Bogers, M., Chiesa, V., and Frattini, F. (2018). Creating and Capturing Value from Big Data: A Multiple-Case Study Analysis of Provider Companies. Technovation. 84-85. 21-36. 10.1016/j.technovation.2018.07.004.

van de Ven, M., Lara Machado, P., Athanasopoulou, A., Aysolmaz, B. and Turetken, O. (2023) Key performance indicators for business models: a systematic review and catalog. *Inf Syst E-Bus Manage* 21, 753–794. https://doi.org/10.1007/s10257-023-00650-2

van Veenstra, A. F. and van den Broek, T. (2015). A Community-driven Open Data Lifecycle Model Based on Literature and Practice. 183-198. 10.1007/978-3-319-08081-9_11.

Voigt, P. and Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide. 10.1007/978-3-319-57959-7.

Wahyuni, D., 2012. The research design maze: Understanding paradigms, cases, methods and methodologies. *Journal of applied management accounting research, 10*(1), pp.69-80.

Weber M., Žarko I. (2019) A Regulatory View on Smart City Services. *Sensors.*19(2):415. https://doi.org/10.3390/s19020415

Weber, T. and Kranzlmüller, D. (2019) "Methods to Evaluate Lifecycle Models for Research Data Management" *Bibliothek Forschung und Praxis*, vol. 43, no. 1, 2019, pp. 75-81. https://doi.org/10.1515/bfp-2019-2016

Webster, J. and Watson, R. T. (2002) 'Analyzing the past to prepare for the future: writing a literature review', *MIS Quarterly*, 26(2).

Yin, R. K. (2014). Case Study Research Design and Methods. Sage.

Zachman, J. A. (1987). A framework for information systems architecture. IBM systems journal, 26(3):276–292.

Appendix A – Participant Consent Form

Consent Form

I.………………………………………agree to participate in Claudia Roessing's research study titled "Modelling Variations in Data Lifecycles – demonstrated in a Smart City Context"

Please tick each statement below:

The purpose and nature of the study has been explained to me verbally & in writing. I've been able to ask questions, which were answered satisfactorily. ☐

I am participating voluntarily. ☐

I give permission for my interview with Claudia Roessing to be video/audio recorded. ☐

I understand that I can withdraw from the study, without repercussions, at any time, whether that is before it starts or while I am participating. ☐

I understand that I can withdraw permission to use the data right up to publication/anonymization/submission of thesis until August 2023. ☐

It has been explained to me how my data will be managed and that I may access it on request. ☐

I understand the limits of confidentiality as described in the information sheet ☐

I understand that my data, in an anonymous format, may be used in further research projects and any subsequent publications if I give permission below: ☐

I agree to quotation/publication of extracts from my interview/focus group. ☐

I do not agree to quotation/publication of extracts from my interview. ☐

I agree for my data to be used for further research projects. ☐

I do not agree for my data to be used for further research projects. ☐

I agree to share my email address with the researcher if I want to have a copy of data that I provided. ☐

Signed……………………………………. Date……………….

Participant Name in block capitals …………………………………………...

I the undersigned have taken the time to fully explain to the above participant the nature and purpose of this study in a manner that they could understand. I have explained the risks involved as well as the possible benefits. I have invited them to ask questions on any aspect of the study that concerned them.

Signed……………………………………. Date……………….

Researcher Name in block capitals …………………………………………...

If during your participation in this study you feel the information and guidelines that you were given have been neglected or disregarded in any way, or if you are unhappy about the process, please contact the Secretary of the Maynooth University Ethics Committee at research.ethics@mu.ie or +353 (0)1 708 6019. Please be assured that your concerns will be dealt with in a sensitive manner.

For your information the Data Controller for this research project is Maynooth University, Maynooth, Co. Kildare. Maynooth University Data Protection officer is Ann McKeon in Humanity house, room 17, who can be contacted at ann.mckeon@mu.ie. Maynooth University Data Privacy policies can be found at https://www.maynoothuniversity.ie/data-protection.

Two copies to be made: 1 for participant, 1 for PI

Appendix B – Metamodel with Design Requirements

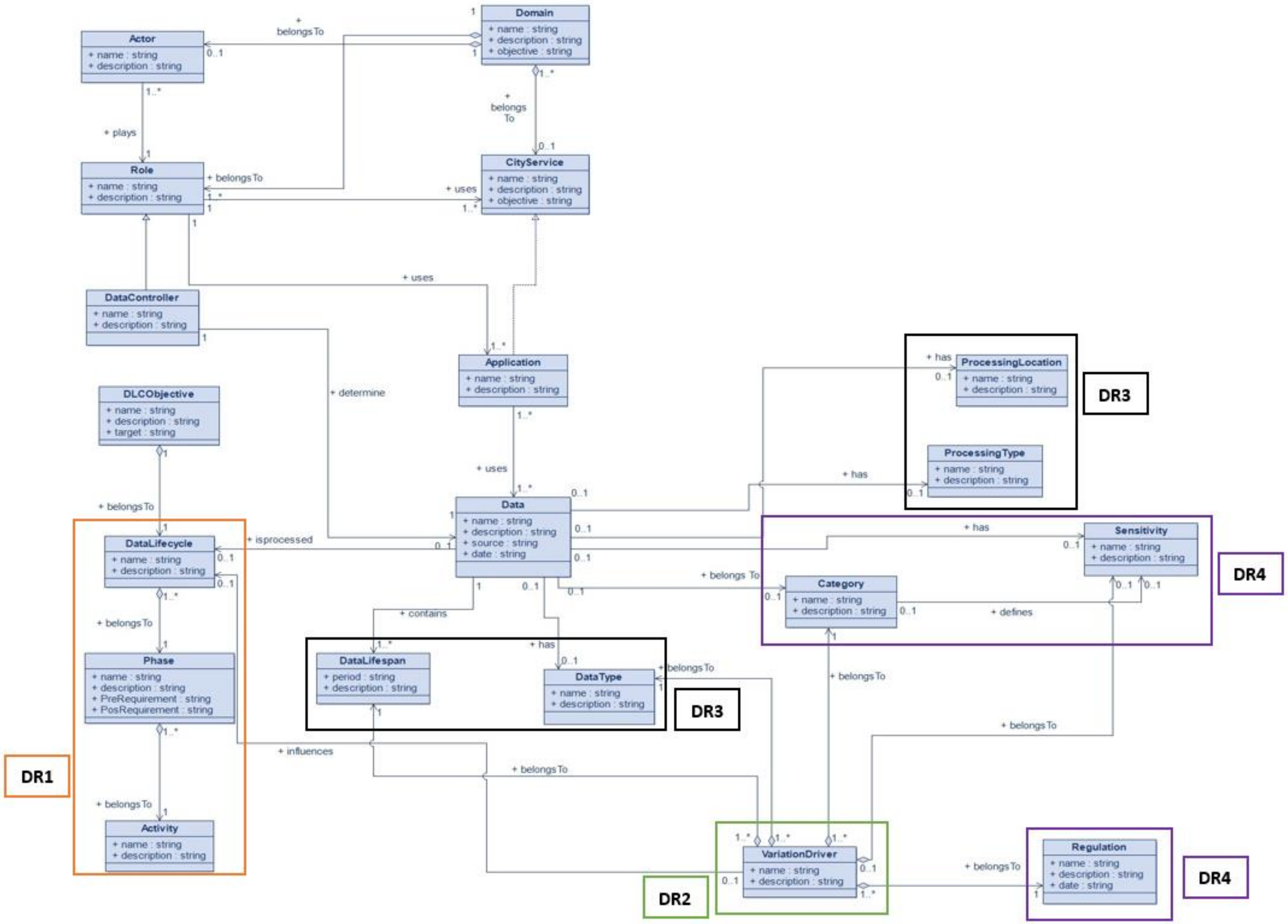

Figure 48 Metamodel with Design Requirements

Appendix C – Interview Guide – Problem Identification

What are the main challenges and issues faced to manage data?

Does the city council use a specific data management tool, e.g. data lifecycle? If yes, how did the council choose the framework? Do the phases change because of different data management requirements?

Does the city council use Enterprise Architecture?

Is there a pre-determined list with all the inputs, outputs, and quality control from each stage of a lifecycle?

Is there quality control in all stages of the data management process?

How the city council ensures data compliance in processing data?

Based on your experience, what are the relevant components of a data lifecycle?

Does the city council support the reuse of data?

Any data is shared/processed or stored by third parties on behalf of the city council?

For how long data is kept by the city council or third party?

Is data deleted by the city council?

Does the city council get permission from citizens to process their personal data?

Can a citizen ask the city council to check his/her data and ask to delete it?

Do you know the information flow from collection until the service is provided to the citizens?

Is there anything would you like to add?

Appendix D – Interview Guide - Evaluation

Do you think the model is relevant based on data that is processed in your city council?

Do you think the concepts, attributes and relationships are adequately representing the process?

Are all the concepts of the model relevant for addressing the challenges that your city council faces?

Do you think the additional concepts are useful to better understand the data management process? Are they helpful to make decisions?

Is it easy to understand and comprehend the model?

Does the model (concepts) provide the complete representation of the data management process?

Can you think of any missing concepts in the model?

How would you used the proposed model?

Do you think the proposed model is more effective than the current one that you are using?

Do you think the proposed model would provide a better outcome compared to what you are using now?