Integrated topological representation of multi-scale utility resource networks

Thomas Tavernor Gilbert

School of Engineering

Newcastle University

Thesis submitted for the degree of Doctor of Philosophy

December 2021

Abstract

The growth of urban areas and their resource consumption presents a significant global challenge. Existing utility resource supply systems are unresponsive, unreliable and costly. There is a need to improve the configuration and management of the infrastructure networks that carry these resources from source to consumer and this is best performed through analysis of multi-scale, integrated digital representations. However, the real-world networks are represented across different datasets that are underpinned by different data standards, practices and assumptions, and are thus challenging to integrate.

Existing integration methods focus predominantly on achieving maximum information retention through complex schema mappings and the development of new data standards, and there is strong emphasis on reconciling differences in geometries. However, network topology is of greatest importance for the analysis of utility networks and simulation of utility resource flows because it is a representation of functional connectivity, and the derivation of this topology does not require the preservation of full information detail. The most pressing challenge is asserting the connectivity *between* the datasets that each represent subnetworks of the entire end-to-end network system.

This project presents an approach to integration that makes use of abstracted digital representations of electricity and water networks to infer inter-dataset network connectivity, exploring what can be achieved by exploiting commonalities between existing datasets and data standards to overcome their otherwise inhibiting disparities. The developed methods rely on the use of graph representations, heuristics and spatial inference, and the results are assessed using surveying techniques and statistical analysis of uncertainties. An algorithm developed for water networks was able to correctly infer a building connection that was absent from source datasets.

The thesis concludes that several of the key use cases for integrated topological representation of utility networks are partially satisfied through the methods presented, but that some differences in data standardisation and best practice in the GIS and BIM domains prevent full automation. The common and unique identification of real-world objects, agreement on a shared concept vocabulary for the built environment, more accurate positioning of distribution assets, consistent use of (and improved best practice for) georeferencing of BIM models and a standardised numerical expression of data uncertainties are identified as points of development.

Acknowledgements

Supervision team

Philip James	Newcastle University, UK
Stuart Barr	University of Melbourne, Australia
Luke Smith	Newcastle University, UK
Jeremy Morley	Ordnance Survey

Other supporting individuals

Trevor Pearman, James Goodyear, Craig Robson, Neil Harris, Qingyuan Ji, Melissa Ware, Chris Jones, Will Seabury, Carsten Roensdorf, Nick Groome, Richard Bennett, Andrew Davey, Anne Spaa

Sponsors

Engineering and Physical Sciences Research Council

Ordnance Survey

Impact statement

The following journal articles are outputs of this PhD project. I have based Chapter 3 on the content of the IJGI paper from 2018 and the first half of Chapter 4 (primarily section 4.2) on the content of the CEUS paper from 2021.

> **Gilbert, T**; James, P; Smith, L; Barr, S; Morley, J. *Topological integration of BIM and geospatial water utility networks across the building envelope*. Elsevier Computers, Environment and Urban Systems, Volume 86, 2021, 101570, ISSN 0198-9715, https://doi.org/10.1016/j.compenvurbsys.2020.101570.

> **Gilbert, T**; Barr S; James, P; Morley, J; Ji, Q. *Software systems approach to multi-scale GIS-BIM utility infrastructure network integration and resource flow simulation*. ISPRS Int. J. Geo Inf. 7 (8) (2018) 310, https://doi.org/10.3390/ ijgi7080310.

The following discussion paper was written as a collaboration between members of the Integrated Digital Built Environment (IDBE) working group, Open Geospatial Consortium, buildingSMART International and Ordnance Survey. My contribution was as part of a remote working placement with Ordnance Survey and as a guest member of IDBE. Although I am the lead author of the document, the content comprises contributions from a wide range of individuals and represents an international consensus. It is cited extensively throughout this thesis and my contribution is an output of this PhD project.

> **Gilbert, T**; Roensdorf, C; Plume, J; Simmons, S; Nisbet, N; Gruler, H-C; Kolbe, TH; van Berlo, L; Mercer, A. (2020). *Built environment data standards and their integration: An analysis of IFC, CityGML and LandInfra*. https://portal.ogc.org/files/?artifact_id=96354

Prior to publication of the above paper, I was invited to speak at GeoBIM 2019 in Amsterdam; we used feedback from the audience to refine the content. Following publication of the paper, I was asked to speak at both IDBE webinars that we ran in April 2020, providing introductory context to the other presentations. Recordings of these webinars can be found at the following links:

> "Bringing AEC Information into a Geospatial Context": https://vimeo.com/412669417
> "Modelling Resilience through Digital Twinning": https://vimeo.com/412623790

The following conference paper is a collaboration between the University of Applied Sciences HFT Stuttgart and the Centre for Energy Systems Integration (CESI) at Newcastle University. I was asked to contribute through discussions of the research, critiquing the paper and writing some of the literature review.

Zirak, M; Royapoor, M; **Gilbert, T**; *Cross-platform energy modeling for scalable urban energy simulation: a case-study*. IAPE, 19, 2019.

I have presented twice at GISRUK – in 2018 and 2020. The 2018 GISRUK paper was an early incarnation of the IJGI journal paper (mentioned earlier) and so is not detailed here. The 2020 GISRUK paper concerns research and development on a geospatial speech recognition and web visualisation technology for which I interrupted my PhD and secured funding for approximately two months of work – that research is entirely independent of the PhD and hence is also not detailed here.

In place of an appendix, the computer code that I have written for implementation of the methods of this thesis and the journal papers are provided in the below Git repositories. Further information on the technical implementation is provided within the thesis.

https://github.com/ttgilbert/geobim
https://github.com/ttgilbert/sensor-data-processing
https://github.com/ttgilbert/sam-streaming
https://github.com/ttgilbert/uo-websocket-streaming
https://github.com/ttgilbert/web-visualisation

I have chosen to write the remainder of this thesis in the passive voice.

Table of contents

List of figures

List of tables

Acronyms

ADE: Application Domain Extension (context of CityGML)

AEC: Architecture, Engineering & Construction

AlgC: Algebraic connectivity (context of graph theory)

BIM: Building Information Modelling

bSI: buildingSMART International (standards organisation)

BST: British Summer Time

CAD: Computer Aided Design

CAT5: Fluid category 5

CDBB: Centre for Digital Built Britain

CESI: Centre for Energy Systems Integration (USB, Helix site, Newcastle University)

CRS: Coordinate Reference System

CUUID: Common, universally unique identifier

DB: Database

DLT: Distributed Ledger Technology

DMA: District Metering Area (related to WDNs)

DT: Digital Twin

DTM: Digital Terrain Model

EPSG: European Petroleum Survey Group

FDC: Frederick Douglass Centre (Helix site, Newcastle University)

GIS: Geospatial Information Science/Systems

GML: Geography Markup Language

GNSS: Global Navigation Satellite System

GUID: Globally Unique Identifier

IFC: Industry Foundation Classes (BIM standard)

INSPIRE: Infrastructure for Spatial Information in Europe

IoT: Internet of Things

ISO: International Standards Organisation

ITN: Integrated Transport Network

JS: JavaScript (programming language)

JSON: JavaScript Object Notation (encoding language)

LiDAR: Light Detection And Ranging (remote sensing technique)

MCDA: Multi-Criteria Decision Analysis

MEP: Mechanical, Electrical & Plumbing (context of BIM)

NICD: National Innovation Centre for Data (Catalyst, Helix site, Newcastle University)

NICA: National Innovation Centre for Ageing (Catalyst, Helix site, Newcastle University)

NWL: Northumbrian Water Ltd. (owns and manages the WDNs in Northeast England)

OGC: Open Geospatial Consortium (standards organisation)

OS: Ordnance Survey (UK's national mapping agency)

PAS: Publicly Accessible Specification

RMSE: Route Mean Square Error

SuDS: Sustainable drainage systems

TOID: Topographic Identifier (OS trademark)

UK: United Kingdom

UO: Urban Observatory (Helix site, Newcastle University)

USB: Urban Sciences Building (Helix site, Newcastle University)

UUID: Universally Unique Identifier

VML: VectorMap Local (OS dataset)

WDN: Water distribution network

WNP: Water network partitioning

WSM: Weighted sum mode

Chapter 1 Introduction

1.1 Cities, smart cities and digital twins

The United Nations has assessed the growth of cities and their resource consumption to be "…the greatest challenge to mankind since we became social" (British Standards Institution, 2014a). Since 2008, most of the world's population lives in cities and the global population is forecast to exceed nine billion by 2050 (British Standards Institution, 2014a, 2014b). Cities comprise a multitude of components and systems such as buildings and infrastructure networks. Existing infrastructure within cities is often unresponsive and costly to maintain, and the UK Government has recognised the need to replace them with "innovative delivery systems to more effectively manage and control resource use in the built environment" (British Standards Institution, 2014a). The interdependencies of such systems and the influence of environmental and human behavioural factors cause them to exhibit non-random, irregular and time-dependent characteristics that distinguish them as complex (Boccaletti *et al.*, 2006; Giudicianni *et al.*, 2018; Saleh, Esa and Mohamed, 2018). The study of complex systems concerns understanding indirect effects (Batty and Hudson-Smith, 2006) by modelling them as a sufficiently representative, manageable number of understood parameters (British Standards Institution, 2014a) and this approach can be applied to the built environment.

The concepts of smart cities and digital twins are commonly used in descriptions of digital representation of the built environment. The terms 'smart city' and 'smart city model' are defined inconsistently but used freely by industry and government (UK BIS, 2013). Whereas a smart city may be considered an urban area that uses data and technology in a coordinated manner (Bari, 2015), it has also been defined as the "effective integration of physical, digital and human systems in the built environment to deliver a sustainable, prosperous and inclusive future for its citizens" (British Standards Institution, 2017). A smart city *model*, however, is clearly a digital representation rather than something real-world. Digital twinning is a similar concept but is not exclusive to urban modelling; in fact, it was first introduced in 2002 for project lifecycle management (Grieves, 2019). In the context of the built environment, the UK's Centre for Digital Built Britain (CDBB) defines a Digital Twin (DT) as a "realistic digital representation of assets, processes and systems" (Centre for Digital Built Britain, 2020). In 2013, the UK Government published an industrial strategy, recognising that "Applying new technology will be a key part of the burgeoning Smart City agenda, where the global market for integrated city systems is set to be worth £200 billion per annum by 2030”. Digital representations of built environments continue to attract significant attention from international standards organisations (Open Geospatial Consortium, 2015; Gilbert *et al.*, 2020) and at political levels in the UK (Cabinet Office, 2018; Geospatial Commission, 2020a, 2020b).

An important factor that challenges the realisation of functional digital representations of built environments is the complex interplay of diverse environments and themes that span a breadth of spatial scales. Forming key components of these building environments are the utility networks that transport the resources of (amongst others) electricity, water and gas to consumers for a range of residential, public and commercial purposes. The physical objects or assets that comprise these networks feature indoors, outdoors, above-ground and below-ground, intersecting the remits of both the Geospatial Information Science (GIS[1]) and Building Information Modelling (BIM) domains.

1.2 Geospatial and BIM domains

Traditionally, the GIS and BIM domains have been regarded as distinct but the boundary between them is becoming blurred (Gilbert *et al.*, 2020). Coarser[2] resolution data describing existing environments are generally handled by GIS practitioners, while the finer resolution designs of the architecture, engineering and construction (AEC) domain are created by BIM specialists. There is now an accelerating demand for digital representations of our entire built environment, which requires the simultaneous use of data from both domains and this can only be satisfied through greater software interoperability and data integration (Gilbert *et al.*, 2020). Integrated modelling of geospatial and building information is a significant challenge to the development of the spatial data infrastructures (Isikdag and S Zlatanova, 2009; Deng, Cheng and Anumba, 2016) that are necessary for the simulation, analysis and visualisation tasks encountered in civil systems engineering (El-Mekawy, 2010; Amirebrahimi *et al.*, 2015a; Borrmann *et al.*, 2015). Although there has been significant research targeting the integration of datasets with each of the two domains (Isikdag and S Zlatanova, 2009; El Meouche, Rezoug

[1] When used as a reference to the professional/practical domain, 'GIS' and 'geospatial' are sometimes used interchangeably in this thesis.
[2] In this thesis, the words coarse and fine are used to distinguish between spatial detail: fine resolution (or granularity) suggests that relatively high precision has been used to represent detail over relatively small distances, areas or volumes; coarseness suggests the opposite.

and Hijazi, 2013; Fosu *et al.*, 2015), there is little evidence of research that focusses on the topologies of utility resource networks than span both domains.

Currently, the real-world continuity of utility resource flow from supply to demand is belied by the discontinuity of digital representations. The diversity of themes and professional practices encompassed by the relevant scales has given rise to data standards that are heterogeneous and incongruous, and thus instances of their schemas that are difficult to integrate. The representation of utility networks as manageable end-to-end systems demands a focus on integrating the topologies of currently digitally unintegrated subnetworks. In support of this, there is a need to identify the extent to which existing representations may be readily utilised to this end and the further development of data standards that is needed to address deficiencies.

This thesis focusses on the challenges facing methods of integrating the topologies of utility networks on the scale at the GIS and BIM domain interface – the building envelope – with a view to guiding the development of the underpinning data standards. Before looking more closely at the data standards and existing integration methods, it is important to clarify and concretise the value of this effort. The following section outlines some broad challenges and describes three use cases that direct the subsequent research of this thesis.

1.3 Utility infrastructure challenges and use cases

The sustained and sustainable provision of resources to inhabitants of urban areas is critical and the digital representation of the networks that transport these resources is fundamental to their effective management. There is a growing recognition of the need to enable more informed decision-making and the necessity of achieving integration of digital representations of complex natural and built environments (Bolton *et al.*, 2018; Hetherington and West, 2020). The design and engineering of utility infrastructure is a complex task due to a multitude of constraints, the necessary involvement of diverse domain specialists and the need to represent information on multiple spatial scales (Borrmann *et al.*, 2015). Both electrical power grids and water distributions networks are examples of complex utility networks that are constrained by their geographic setting, comprising multiple interconnected and interacting parts (Boccaletti *et al.*, 2006; Yazdani and Jeffrey, 2011) that span nearly every terrestrial spatial scale. Additional complexity and challenge is added to the modelling and analysis of these network by their interdependencies (Solomakhina *et al.*, 2015), ageing of physical assets (Tang, Parsons and Jude, 2019; Ahopelto and Vahala, 2020) and that these assets are often located underground

(Geospatial Commission, 2019, 2020a). The performance of utility systems depends on their network topologies (Simone *et al.*, 2018) – this concerns the connectivity at a particular spatial scale, across multiple scales and between different utilities. An ability to accurately measure and improve performance through maintenance, reconfiguration and modification depends on integrated representations of these topologies, which demands an ability to integrate data from both the geospatial and GIS and BIM domains.

There are multiple use cases of relevance to the domain of GIS-BIM integration, many of which are summarised by Liu *et al.* (2017); for some of the use cases, there is a need to integrate data describing utility resource infrastructure and a further subset require, in particular, a representation of the connectivity of such infrastructure. The outer shell of a building – also referred to as the building envelope – is approximately at the spatial boundary of the GIS and BIM domains. In this thesis, three use cases requiring the connectivity of utility networks across the building envelope are identified for research focus: electricity demand-side management, water network partitioning, and spatial-topological configuration planning. Peak electricity loading can be reduced through real-time-pricing incentives, minimising consumer costs and outage risks, and this is supported by the ability to disaggregate and trace demand through the finer scales of urban areas. Dynamically configurable water network topologies have been proposed for both leakage detection capabilities afforded by partitioning and the connection redundancy of large-scale looped networks. The planning of utility services is constrained by spatial factors such as physical obstruction and connection distances, which influence the trade-off between minimising engineering cost and maximising topological redundancy. These use cases require the accurate integration of the topologies of the finer scale internal building consumption networks with the coarser scale distribution networks of urban areas.

1.4 Research questions, aims and objectives

Chapter 2 considers the above use cases in more detail and then provides an analysis of the key built environment data standards and a critique of existing methods of their integration, from which research gaps are identified.

Research questions were derived through an iterative process of prototyping, testing, demonstration, feedback, discussion and refinement. Face-to-face meetings, video calls, email and accompaniment of engineers during on-site visits were used to engage multiple representatives of stake-holding organisations in this iterative process of converging on an appropriate research questions – these interactions are detailed further in the case studies of Chapter 3 and Chapter 4. Those whose discussion influenced the direction of the research include: members of the Estates department at Newcastle University; the BIM manager Bowmer and Kirkland (the prime contractor of the Helix project - see Chapter 3); the Research & Development Manager and Continuous Improvement Lead at Northumbrian Water (see Chapter 4); the Senior Project Manager at NG Bailey (the engineering contractor for Helix); a BIM researcher at Northumbria University; academics from Newcastle University, TU Delft and TU Munich; and employees of BuildingSMART International, Open Geospatial Consortium and Ordnance Survey.

The research gaps are identified in section 2.5 as appropriate levels of information abstraction, requirements and methods for identification of real-world objects, and standardised representations of location uncertainties in support of integrating utility networks topology. Through the research design process, the creation of a new data standard to support the integration of utility network topologies was deemed to be unnecessary and unlikely to gain traction with implementers. Instead, the research questions were chosen to address how existing data standards can be used effectively in their existing form and how they can be modified to improve capability.

1. How can existing datasets be leveraged to construct digital representations of utility network topologies across the building envelope and how does this support the priority use cases?
2. How can existing data standards be modified to support the priority use cases where existing data cannot be leveraged to sufficient effect?

The aim of this research is to address the research questions by devising and prototyping integration methods that elicit an understanding of how existing datasets and data standards can be leveraged for construction of utility network topologies across the building envelope. Section 2.6 describes a research approach that is based on the ideas of pragmatism and heuristics in engineering, with case studies used to better understand use cases, which are satisfied by methods and their underpinning techniques. The research goal is to develop methods of asserting functional relationships between disjoint and disparate digital representations of utility networks. The objectives are as follows:

1. Develop a set of priority use cases to understand the need for integrated digital representation of multi-scale utility network topologies.

2. Carry out a review of the domains of 3D Urban GIS and BIM with a focus on the representation of utility networks at and around the scale of the building envelope.
3. Examine the relevant data standards in these two domains and identify the key disparities that present a challenge to the integration of utility network datasets.
4. Analyse and critique existing methods of urban data integration, focussing on weaknesses in their applicability to utility network topology[3], and identify research gaps.
5. Address the research gaps through the exploration of case studies.
6. Design and prototype methods that satisfy the general requirements the use cases through addressing specific requirements of case studies.
7. Discuss the findings from the case studies in the context of the use cases and data standards.
8. Highlight areas of beneficial future research.

1.5 Thesis structure

The remainder of this thesis addresses the aim and objectives over several chapters. Chapter 2 contains the literature review and research approach; it considers the geospatial and BIM domains, the data standards in each that dominate urban environment modelling, the disparities between these standards and the challenges these disparities present to the integration of utility network topology, the methods that have been devised to integrate instances of these standards, and the gaps in research that merit further exploration; the chapter culminates by defining a research approach. The subsequent two chapters detail the software prototyping aimed at enabling real-time visualisation of electricity flows (Chapter 3), the topological placement of consumer connection points in a potable water distribution network (Chapter 4) and the optimisation of spatial-topological utility network configuration (also Chapter 4). Chapter 5 discusses the methods and results of Chapter 3 and Chapter 4 in the context of the literature review (Chapter 2), identifying integration potential given existing data representations and opportunities for development of the standards where existing characteristics inhibit integration as required by use cases. Chapter 6 concludes the thesis.

1.6 Summary

Cities and other urban areas are becoming more densely populated but ageing infrastructure is often unable to cope. The consequent of this is an increasing demand on resources. The complex

[3] Network topology concerns the connectivity of components of a network and is defined more thorough in section 2.2.1.

interplay of the various components of built environments needs to be represented digitally and integrated in order that resource delivery systems can address this problem through analysis and optimisation. The relevant components of built environments span the spatial scales of geospatial and BIM domains, which exhibit disparities that inhibit integration of their disjoint datasets. There are several use cases that require the integration of the network topology of utility infrastructure across the building envelope, three of which are discussed in the next chapter as foci for the remainder of the thesis.

Chapter 2 Literature review and research approach

2.1 Introduction to the literature review

This thesis is multidisciplinary, covering a breadth of domains and demanding a consideration of multiple topics. The themes that need to be covered by this literature review are separated into three main sections: use cases for utility network topology integration (2.2), the domains and data standards of relevance to this integration (2.3) and existing methods of integration (2.4). The premise of this order is that there should be a well justified reason for integrating datasets (provided by use cases) before considering *how* integration is or should be performed (existing methods). A review of the domains and data standards is placed before that of existing methods because the domains and standards are fundamental to the methods that have been developed. Section 2.4 analyses and critiques existing methods throughout and culminates in a synthesising critique (2.4.8). Research gaps deriving from this critique are identified in section 2.5, and section 2.6 presents the approach that this thesis takes to addressing these gaps. The entire chapter is summarised in section 2.7.

2.2 Use cases

2.2.1 Introduction and use case selection

"…if an unquestionable objective is to minimise the intervention required from a user to achieve a viable and useful level of integration, there is a pressing need to articulate both existing and plausible future use cases against which any approach can be implemented and assessed" (Gilbert *et al.*, 2020)[4]. The previous chapter highlighted a set of three use cases on which this section now elaborates.

Use cases may be understood as informal scenarios that depict the behaviour of a system with respect to the needs of a user (Amirebrahimi *et al.*, 2015b). They provide guidance that is critical

[4] This discussion paper is cited and quoted extensively throughout this chapter and the remainder of this thesis. The paper was written in collaboration with Ordnance Survey, Open Geospatial Consortium and buildingSMART International through the Integrated Digital Built Environment working group; the author of this thesis is the main author of the cited paper – this authorship is an output of the PhD research project.

to the design, development and evaluation of technical approaches and there are myriad examples for urban data integration (Liu *et al.*, 2017). In the context of this thesis and research, a use case is understood to be a generalised example rather than a specific, concrete example, which is understood as a case study (see section 2.6 for more detail, including a diagram). The use cases are used to justify and frame the research questions and objectives; the case studies are used to test the suitability of the methods at addressing the use cases. This thesis is concerned with addressing the challenges of relevance to the integration of the *network topology*[5] of utility networks – specifically, the functional relationships between components or their arrangement in a network. The research of this thesis is focussed strongly on investigating the suitability of existing digital representations of utility networks for the key uses cases that demand a digital representation of utility network topology. In support of such investigation, the key use cases need to be understood in detail.

The literature review (Chapter 2) and interactions with the stakeholders (listed in section 1.4) were used to identify use cases that could guide this research. Under broad themes of interest to industry (such as planning and maintenance), specific technical requirements were identified. Many of these requirements demand the integration of BIM and geospatial datasets; these became the use cases of initial consideration. From this initial set, those that did not require the representation of network connectivity were eliminated. The strongest examples of the remaining within-scope use cases were selected for focus.

The planning of a construction project might require the detailed spatial representation of both an architectural BIM model and a city model in order to present a 3D rendering to a client interested in the outputs of a shadow or field-of-view analysts. However, such 3D walk-through use cases were eliminated because there is no requirement for the representation of any connectivity. Similarly, clash detection between the foundation of a proposed new-build and existing underground pipes depends on spatial data alone and would not be within scope (although the research of Chapter 4 does later demonstrates how heuristics can be used to infer connectivity and the presence of unrepresented assets that could present a clash risk). Use cases involving quantity take-offs or financial calculations were also not considered – the ability to tabulate values across an integrated, searchable knowledgebase of an urban environment does not necessarily depend on any network connectivity, even if the assets of interest are those from a utility network. The use cases could be categorised into those that demand a representation of

[5] Network topology is distinct from spatial topology, which is described in footnote 22.

the urban environment that is predominantly spatial, tabular or topological; only those that depend ultimately on a topological representation of utility networks were selected for this research.

Three use cases were chosen for focus: the visualisation of existing supply and usage within a network (focussing on electricity), the partitioning of water distributions networks (WDNs) and the spatial-topological configuration planning of utility networks more generally. The following subsections detail each use case and explain the importance of integrated utility network topology modelling for each.

2.2.2 Electricity demand side management and dynamic pricing

The provision of electricity has become essential to modern life and its seamless supply has been cited as a requirement for industrial growth and increases to quality of life[6] (Srinivasan *et al.*, 2017); it also presents a major challenge to today's society in which economic, ecological and political concerns are at stake (Aussel *et al.*, 2020). Electricity demand and supply both fluctuate in ways that are not always predicted (UK Parliamentary Office of Science & Technology, 2014), especially given the intermittency of increasingly popular renewable resources (Bu, Yu and Liu, 2011). Storage of electricity is generally not cost-effective such that its generation must be concurrent with demand (Koomey and Brown, 2002). Due to the variability of unregulated demand and the need for generation to match peak demand (supply is usually inflexible), much of this capacity is above average demand and thus unused, constituting an inefficiency and lost opportunity (Dutta and Mitra, 2017). However, high peak demand risks damage to supply systems (Moholkar, Klinkhachorn and Feliachi, 2004) and the size of the peak determines network charges to consumers (Sun, Wang and Huang, 2010; UK Parliamentary Office of Science & Technology, 2014).

Demand side management (DSM) can be used to reduce risk and cost. The principle of DSM is that demand can be fitted to production rather than vice versa (Aussel *et al.*, 2020). This management can be categorised into load reducing and load shifting strategies (Mohsenian-Rad *et al.*, 2010), along with storage of energy[7] for later use (Koomey and Brown, 2002; Tang, Wang and Li, 2019). While remote management of end-user consumption by load aggregators can achieve lower system stress and improve efficiency of electricity distribution (Baccino *et al.*, 2013; Saleh, Pijnenburg and Castillo-Guerra, 2017), active demand methods can achieve

[6] This is clearly subjective.
[7] Such as via chilled water or ice storage tanks (Sun, Wang and Huang, 2010).

similar outcomes by placing onus the user to alter their behaviour: price reduction is used as an incentive for reduction of consumption during peak hours (Ma *et al.*, 2016; Ahmed *et al.*, 2018; Aussel *et al.*, 2020) and dynamic pricing is an emerging DSM technique[8] that can reduce peak load by varying prices according to demand (Dutta and Mitra, 2017; Ahmed *et al.*, 2018) – it has been shown to be economically and environmentally advantageous (Leon-garcia, 2010; Desai and Dutta, 2013; Finn and Fitzpatrick, 2014). For example, dynamic pricing can be used to vary building zone temperature set-points to minimise peak cooling demand (Lee and Braun, 2008; Tang, Wang and Shan, 2018). Real-time pricing (RTP) is a DSM method in which prices vary at regular, frequent intervals (as small as sub-hour), increasing the efficiency of the pricing scheme by reflecting the demand-supply balance in real-time[9] (Moholkar, Klinkhachorn and Feliachi, 2004; Dutta and Mitra, 2017). A consumer's demand requirement can also vary with time (Bu, Yu and Liu, 2011) and RTP programmes are based on the willingness of consumers to exploit this in response to fluctuating prices (Siano, 2014) for their own monetary savings while simultaneously minimising the supplier's risk of outages.

Utilities are concerned not only with outages on a system-wide scale due to total system demand; local outages are more common than system outages and are extremely costly if they interrupt economic activity (Koomey and Brown, 2002). DSM that reduces peak loading on the finer spatial scales of local urban areas has the potential to reduce outages and save money. However, such schemes require suitable technology to communicate and manage the frequent changes (Baccino *et al.*, 2013; Dutta and Mitra, 2017) and an integrated representation of the topology of the networks across all of the scales of the urban area. For the owner of a campus or estate to be able to manage their total energy demand (for example, to suppress the load during known peak hours or in response to an unexpected real-time price increase), a disaggregation of demand down to each building – and then inside it – would allow them to trace through to the type and location of the consumption units that are incurring the highest load (for example, a set of computers in a particular room) and then propose a load reduction strategy where they assess there to be flexibility. From another perspective, a conscientious user may wish to reduce their impact on peak demand (perhaps for ecological reasons) and a reduced-complexity visualisation that shows their contribution to the total demand would enable them to shift their behaviour quickly and in an informed manner – for example, they could decide to reschedule computer updates to take place during times of lower demand. This

[8] Other techniques include time of use (TOU) and critical peak pricing (CPP) (Ahmed *et al.*, 2018)
[9] Or near-real-time, depending on the threshold frequency of updates for classification as 'real-time'.

integrated representation of urban topology enables a harnessing of flexibility in demand to counter the inherent inflexibility of supply.

2.2.3 Water network partitioning

Water distribution networks (WDNs) are critical infrastructure, providing the clean water needed for socio-economic prosperity and population health (Matthews, 2016; Meng *et al.*, 2018; Giudicianni, Herrera, di Nardo, Greco, *et al.*, 2020). A significant portion of water infrastructure in the UK is deteriorating, resulting in background leakage and bursts that disrupt supply, waste water, cause further damage and are costly to fix (Tang, Parsons and Jude, 2019; Ahopelto and Vahala, 2020). In England and Wales, approximately 3.17 billion litres of water (21% of public supply) are lost to leakage every day (Price Waterhouse, 2019) and the global annual cost of lost water is estimated as USD 39 billion (Liemberger & Wyatt, 2019). Due to the wide geospatial distribution of WDNs and their multiple points of access, they are vulnerable to contamination and costly damage (Hart and Murray, 2010). WDNs can consist of thousands of components and the complexity of their interrelationships makes it difficult to predict their performance under various scenarios (Perelman and Ostfeld, 2011). Despite this, demands are being placed on water companies to improve their service against performance indicators such as supply continuity, water discolouration, energy efficiency and sufficient pressure at the point of consumption (Wright *et al.*, 2014). Although insufficient pressure inhibits supply, maintaining operational pressure close to the threshold minimum can reduce water loss and burst frequency (Wright *et al.*, 2014).

Water network partitioning involves the compartmentalisation of water distribution networks (WDNs) into district metering areas (DMAs), a practice that has grown in popularity (Charalambous, 2008). DMAs help to improve pressure management[10], which is the only controllable factor[11] that affects leakage once pipes have been laid (Germanopoulos and Jowitt, 1989), and locate leakages once they do occur (Taillefond and Wolkenhauer, 2002; Wright *et al.*, 2014; Azevedo and Saurin, 2018; Ahopelto and Vahala, 2020; Khoa Bui, S. Marlim and Kang, 2020). Using a comparison between actual monitored net flow rates in DMAs at times of expected minimum activity (usually at night) and flow rates that represent legitimate usage

[10] DMAs can also be used for burst detection (Wu *et al.*, 2016) and load balancing (Ferrari, Savic and Becciu, 2014).

[11] Other factors include "…movement and characteristics of the soil in which the pipes are laid, the degree of deterioration of water mains and pipes, the quality of fittings, materials and workmanship, as well as the possible effect of traffic loading in causing the failure of buried pipelines." (Germanopoulos and Jowitt, 1989)

during those periods, leakage can be estimated (Khoa Bui, S. Marlim and Kang, 2020) and located to a specific area for intervention.

However, the use of fixed DMAs has its disadvantages, including reduced operational flexibility and network resilience (Scarpa, Lobba and Becciu, 2016; Giudicianni, Herrera, di Nardo and Adeyeye, 2020). Resilience is widely understood as the capacity of a system to resist, absorb, withstand and rapidly recover from exceptional conditions (Johansson, 2010; Amarasinghe, 2014; Hosseini, Barker and Ramirez-Marquez, 2016; Butler *et al.*, 2017; Meng *et al.*, 2018) and has been increasingly pursued in the management of WDNs (Wright *et al.*, 2014). Manual valve operations are required during failures, which reduces the natural redundancy in connectivity of these looped networks (Wright *et al.*, 2014). Along with factors such as structural integrity and pumping power supply backups, redundancy is a key aspect[12] of resilience in water infrastructure systems (Matthews, 2016). Dynamically reconfigurable water network topologies have been proposed in place of fixed DMA, allowing for both the leakage detection capabilities of partitioning and the connection redundancy of large-scale looped networks[13]: multifunction network controllers modify the topology and continuously monitor the dynamic hydraulic conditions (Wright *et al.*, 2014), for which implementation methods have been proposed and demonstrated (e.g. Giudicianni, Herrera, di Nardo and Adeyeye, 2020).

However, WDNs do not always represent accurate or complete topologies: the exact points of connection of consumer nodes (such as buildings and other facilities) are not always known or represented digitally; when they are, the finer-scale internal building consumer networks are not integrated with that of the WDN – from the point of view of the utility provider, the topology stops at the point from which the building is supplied. BIM mechanical, electrical and plumbing (MEP) models can represent the spatial layout, network topology and details of consumption units within buildings. In the context of dynamically configurable WDNs, the accurate connection of these more granular subnetworks with WDNs is needed for more precise modelling of the downstream impacts of various scenarios and hence optimisation of real-time topology reconfiguration.

Prioritising continuity of supply to critical facilities is justification for this use case. Burst mains have caused outages to hospitals in the UK (BBC, 2013; Guardian News, 2018). Internal

[12] Other factors include water stagnation in dead-end branch pipes and discoloration due to high spatial and temporal variation in flow rates (Wright *et al.*, 2014).
[13] Leakage rates are also lower in networks with higher redundancy due to lower average zone pressures and less pressure variability (Wright *et al.*, 2014, 2015).

networking data from a BIM model of a hospital, when integrated with that of the WDN, may indicate that this priority consumer is connected to a specific node in the network and any dynamic reconfiguration should not allow a pressure drop at that supply point – beyond the multiple everyday critical functions of a hospital, for example, an internal sprinkler system might depend on this continued supply for time-critical firefighting.

2.2.4 Network configuration planning

Beyond the integrated digital representation of existing networks, there is a future planning use case: existing BIM models can be used for urban planning and existing urban plans can be used for building design; buildings and their surrounding geospatial environment can be designed in the context of each other. The position of a building on a construction site determines its distance to existing building services infrastructure (Peckiene and Ustinovičius, 2017) and spatial obstruction is a significant factor in, for example, pipeline planning processes (Zhao, Liu and Mbachu, 2019). "The presence, layout, and organisation of underground utilities directly affect the value that the land can continuously deliver. It may limit the potential for future development and use or the capacity to host new infrastructure and may present significant obstacles, risks, and nuisances for owners, developers, engineers, and users of the land." (Yan, Van Son and Soon, 2021)

However, there are topological design factors that are influenced by spatial constraints. For WDNs, although backup power and structural stability are perhaps the two most important aspects for reliability[14] of water provision (Matthews, 2016), some topological attributes influence system resilience (Meng *et al.*, 2018) and spatial constraints can impinge on feasibility. For example, a mains water supply point of higher topological redundancy might be situated farther from a building than a point of lower redundancy, or there could be obstruction along a transit to a nearby and topologically favourable supply point. The redundancy reduces the chance of service interruption in the event of, for example, a pipe burst along a supply route. However, an appropriate trade-off has to be found between the cost of engineering and the reliability afforded by a topologically superior configuration (Oliker and Ostfeld, 2013).

This trade-off can be optimised through integrated design and this requires integrated data representations. A water provider can ensure that a particular node of a new WDN layout that has high topological redundancy is well situated for connection to a high-demand, critical

[14] Reliability is considered here to be a result of resilience and robustness (amongst other potential factors).

facility such as (returning to the example used in 2.2.3) a hospital. The engineering required to connect to the hospital could be reduced by placing the connection close to its entry point; conversely, for an existing WDN layout, the internal network of the hospital could be configured to ensure that the facility will draw water from a side of the building that is closest (or less obstructed). Such optimisation can be applied to other spatially constrained critical infrastructure networks for which network robustness can be influenced by topology, such as electrical power systems (Baldick *et al.*, 2009; Rezaei, 2016; Robson, 2016).

The different utilities such as electricity, gas and water are interdependent (coupled) and configuration planning can benefit from accounting for this. Infrastructures interact with each other both due to direct physical connection and spatial proximity (or other spatial relations), and these complex interactions can result in cascading effects in which the failure of one system causes the failure of another dependent system (Becker, Nagel and Kolbe, 2011). Examples of vulnerabilities to such effects include electricity systems that are powered by gas, electricity-driven compressors in a gas system (Erdener *et al.*, 2014), thermoelectric power plants that depend on a water supply for their cooling (MISTRAL, 2020), district cooling systems that are powered by natural gas and water pumps that are powered electrically (Solomakhina *et al.*, 2015) - the reliability of the dependent system could hinge on the redundancy of supply of the requisite resource at a point in the network to which the dependent system can be feasibly connected.

The lack of integration of the finer-scale internal building utility network topology with that of surrounding urban areas hinders network configuration planning in a similar way to how it impacts dynamic network partitioning (section 2.2.3). In order to automate such spatial-topological optimisation (whether considering a utility network as a single-resource or multi-resource interdependent system), utility infrastructure across both the finer and coarser spatial scales needs to be analysed as end-to-end networks, which requires the integration of the topologies of the relevant digital representations.

All three of the use cases described in sections 2.2.2, 2.2.3 and 2.2.4 require integration of datasets that span the internal building and external urban scales. Methods of integrating such datasets are often founded on and constrained by data standards, the characteristics of which are determined by the professional domains in which the standards are developed.

2.3 The domains and data standards

2.3.1 Introduction

In order to satisfy the use cases described in section 2.2, integration of the disjoint digital representations of utility network subsystems must overcome the challenge of contrasting professional domains and data standards. On the finer spatial scales of building internals, architects and engineers use building information modelling (BIM) to model digitally the details of internal building networks; on coarser spatial scales, surveyors and urban planners use GIS to represent the distribution networks that supply the buildings. Each domain has its own cultures, practices, technology and, crucially, data standards (Gilbert *et al.*, 2020) – the agreed rules by which the built environment should be represented digitally. The contrasting backgrounds of the domains has strongly influenced the characteristics of these standards and is responsible for disparities in standardisation that must be overcome by integration methods.

This section describes the background to the professional domains of 3D Urban Geospatial Information Science (GIS) and Building Information Modelling (BIM), and the key data standards that influence the challenge of urban data integration with a focus on utility network topology. 3D Urban GIS is used here to refer to the use of GIS for urban data modelling, in which there is often emphasis on representation in all three spatial dimensions. The 3D Urban GIS and BIM domains overlap significantly in their spatial and thematic scope. The outer shells of buildings are a physical interface between interior and exterior environments and are sometimes referred to collectively as the building envelope. This interface is central to the spatial overlap of the two domains and may be regarded as one of the surfaces at which they meet (see Figure 2.3.1-1).

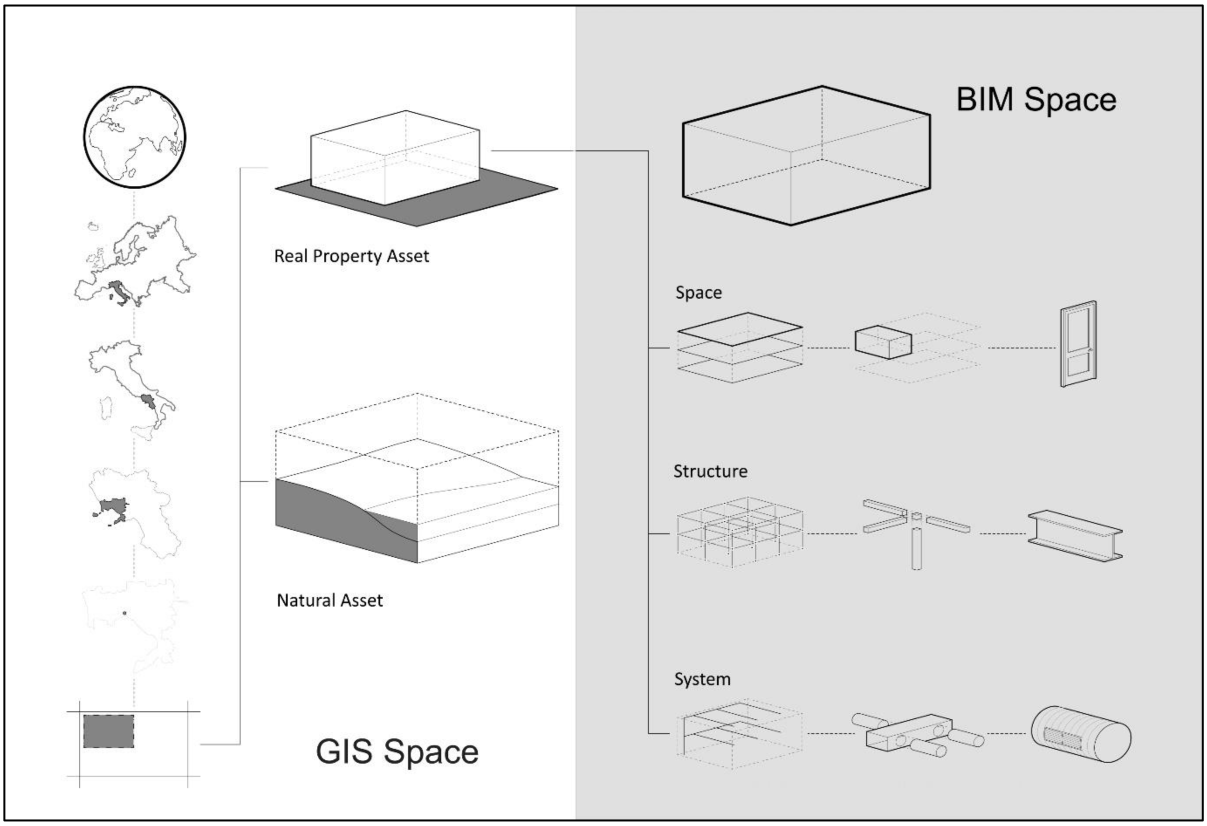

Figure 2.3.1-1 - The environments, spaces, assets and systems that the GIS and BIM domains are normally used to represent, and the interface at which they are considered to meet – the envelope of real property assets or buildings. This is a modified version of a figure by Danelle Briscoe, which features in her book "Beyond BIM: Architecture Information Modeling" (Briscoe, 2016).

The origins and developmental paths taken by the GIS and BIM domains have an influence on today's industry practices and research, the design and evolution of the data standards that are currently used for urban modelling, and the limitations on and opportunities for data integration that this affords. The following two sections describe how the remit of the geospatial domain has evolved to represent objects at this boundary and within the finer scales of building internals, and how the BIM domain grew to represent a broad variety of construction information – beyond the fine details of building internals, through the envelope and out to the coarser scales of urban areas. This provides a context for the subsequent analysis of disparities between the key data standards used in urban modelling. In turn, these disparities influence the design of methods for the topological integration of utility networks whose digital representation is spread across multiple disjoint datasets.

2.3.2 3D Urban Geospatial Information Science and Building Information Modelling

As early as the mid-1990s, computer-generated renderings and walk-throughs were envisaged as techniques to support geospatial design and analysis (Levy, 1995). Around the late 1990s and early 2000s, implementation of these visions was becoming more feasible (Holtier, Steadman and Smith, 2000), with customised GIS software able to represent buildings by defining relationships and interactions between layered 2D polygons in 3D space, retaining information on building form and construction materials. This enabled analyses on energy consumption, noise, lighting, solar photo-voltaic power, occupancy and air pollution (Holtier, Steadman and Smith, 2000) in urban environments. 3D geometric data were also being combined with semantic data to enable more varied analyses on urban scales, such as to support operations for delivery of utility resources (Bernhardsen, 2002). Urban 3D GIS has been fundamental to a broad range of research and applications, including real estate management, environmental simulation, telecommunications and facility management (El-Mekawy, 2010), view quality analysis (Ferreira, Werner and Park, 2015), visual impact assessment of wind turbines (Wrozynski, Sojka and Pyszny, 2016), urban planning (Kolbe and Groger 2003; Kolbe 2009), disaster management (Kolbe, Groger and Plumer, 2005), navigation within the built environment (Lee, 2004), visualisation and exploration of cityscapes, environmental and training simulations, facility management and homeland security (Kolbe 2009; Laat and Berlo 2011). While GIS has been used for planning of linear transport infrastructure (Farooq *et al.*, 2018) such as roads (Jaro and Iguisi, 2015), there has also been extensive research aimed at improving utility network infrastructure and resource management.

Progress on the modelling of energy and water networks has been significant but limited in spatial scope. There has been a shift in energy infrastructure planning to accommodate fluctuating demand and delivery, decentralised generation, cost-free carriers and renewables (Resch *et al.*, 2014) with increasing efficiency leading to significant cost savings (Bernhardsen, 2002). Research has considered the use of GIS for visualisation of smart transmission grids (Li *et al.*, 2010) and measuring the potential for solar photovoltaic energy generation of urban areas using 3D city models (Romero Rodríguez *et al.*, 2017). For water distribution networks, studies have considered the use of GIS for design and analysis of distribution systems and the modelling of their network topologies (Taher and Labadie, 1996), for combining topography with hydraulics to understand network behaviour and optimise network maintenance (Abdelbaki *et al.*, 2017; Awad, Yassin and Ayad, 2017), and for improving leakage control through optimal selection of valves for closure (Mirshafiei *et al.*, 2019). The interdependencies between the different utility types is also an area of research for optimising operations, analysis of vulnerabilities and simulating cascading failures (Becker, Nagel and Kolbe, 2011; Solomakhina *et al.*, 2015). However, these studies rarely consider in detail the features and functions at the finest of spatial scales despite the relevance of these to many of the problems that the researchers are trying to solve.

Although developments have taken GIS capabilities close to being able to represent features across all spatial scales, progress has generally not encompassed the finest resolution networking inside consumer facilities such as buildings. Despite a trend towards partial coverage by the GIS domain of the internals of buildings with data standards such as CityGML (Becker, Nagel and Kolbe, 2013, 2016; Kutzner and Kolbe, 2016), internal building networks are usually designed by engineers in BIM software and its associated data standards, which have been designed with the representation of even finer demand-side details as a requirement (BuildingSMART and Leibich, 2009; Peters, 2010; Zheng *et al.*, 2017). The GIS domain in isolation is largely blind to these details and thus any end-to-end demand-supply representation is dependent on its integration with data from the BIM domain.

The domain of Building Information Modelling (BIM) emerged from methods used for drawing aircraft and boats that predate computers but underpin much of the field of Computer Aided Design (CAD). CAD was conceived to address the inefficiency of hand-drawing and management of discrepancy issues. Advances in computer graphics and software capabilities allowed visualisation for urban planning and building performance assessments under different orientation, geometry and materials. The representation of other dimensions (such as cost and energy), the distinction between object classes and their instances, greater detail on objects' properties and parametric modelling are factors that gave rise to BIM (Levy, 1995; Weygant, 2011; Barnes and Davies, 2014; Kensek, 2014; Briscoe, 2016). Primarily, BIM models represent buildings and indoor environments, including structures and individual components. Architects, engineers and construction (AEC) contractors have been early adopters of BIM and now dominate the use of its functionality (Volk, Stengel and Schultmann, 2014). The benefits afforded by the modelling and analytical capabilities of BIM are well recognised (Weygant, 2011) and, at its highest level of detail (Level 3), a BIM project will be modelled completely in 3D with full collaboration between all contributing disciplines by use of a single, shared and centralised project model (Barnes and Davies, 2014; NBS, 2017).

Although BIM is a relatively young domain, its merits are gaining recognition even at political levels, emphasising its importance and likely traction over the coming decades. The UK Cabinet Office published a Government Construction Strategy (GCS) in 2011, mandating that publicly-funded projects practice “full collaborative 3D BIM” (BIM Level 2) as a minimum by 2016 (Cabinet Office, 2011; HM Government, 2013; Barnes and Davies, 2014), “…with all project and asset information, documentation and data being electronic…” by the same year (British Standards Institution, 2017). Another GCS from 2016 committed the UK Government to developing, alongside industry, the “…next generation of digital standards to enable BIM Level 3 adoption under the remit of the Digital Built Britain Strategy.”, stating that this “would support a fully integrated and collaborative process” in construction (UK Infrastructure and Projects Authority, 2016). These commitments emphasise the importance that has been placed on BIM by high-level decision-makers and thus the importance that should be placed on enabling the integration of BIM within the context of urban data modelling. The National Digital Twin programme has since published their “…approach to delivering a National Digital Twin for the United Kingdom.” (Centre for Digital Built Britain, 2020) in response to a recommendation by the National Infrastructure Commission in their 'Data for the Public Good' report (National Infrstructure Commission, 2017). However, the planned approach includes objectives regarding data modelling, referencing, sharing and integration that indicate much foundational work remains to be done over the following five to ten years.

Research within the BIM domain has concerned both improvements to technology (Daum and Borrmann, 2014; Johansson, Roupé and Bosch-Sijtsema, 2015; Chen, Chang and Lin, 2016) and its usage in applications such as pedestrian route planning (Whiting and Teller, 2006), environmental simulation and disaster management (Arayici, 2007), immersive virtual environments for fire evacuation simulation (Rüppel, Abolghasemzadeh and Stübbe, 2010), safety management and construction conflicts (Hu and Zhang, 2011), the use of graph theory in facility management for administration of access control (Skandhakumar *et al.*, 2016) and construction project risk (Ding *et al.*, 2016). Developments to capabilities in the representation of building internals (as required by modellers) remain largely the responsibility of software vendors, with Autodesk and its products dominating the market. However, buildingSMART has sought to increase the potential for sharing of project and asset information through development of the Industry Foundation Classes (IFC) open data standard (buildingSMART International, 2020a) (discussed in detail in section 2.3.4). The functionality of proprietary BIM software, state of development of IFC and software support for its features all influence the digital representability of various features of facilities and their constituent parts, including utility networks. Autodesk, for example, introduced MEP modelling to Revit in 2006 (Shackelford, 2017) and an initial 'building services' schema was introduced to IFC in 2000 with an extension in 2003 (Liebich, 2010).

The prescriptive nature of BIM (it primarily describes future constructions), the high richness of detail derived from its technical procedures and the use of local engineering coordinate reference systems (CRSs) often make it unsuitable for representation of coarser geospatial scale features. For these reasons, just as the GIS domain requires BIM data for internal building detail, the BIM domain represents relatively little beyond the building envelope until its integration with GIS datasets. The two domains remain disparate, which poses significant hindrance to developing the multi-scale representations of the built environment that are needed for smart city modelling and digital twinning. "In essence, a building is a component of a larger group of features which is linked by infrastructure and other elements to create a holistic system" (Peters, 2010). The GIS and BIM domains need to be considered holistically in order that the flows of people, goods, services and – of relevance to this thesis – utility resources across the building envelope can be better understood and simulated digitally. The following section further details the disparities between the GIS and BIM domains in the context of their intersection and the requirement for their integration, before then analysing the predominant disparities and similarity between two of the key standards – CityGML and IFC – as proxies for built environment data standards more generally.

2.3.3 The challenges of GIS-BIM integration

The realisation of functional digital representations of the built environment depends on integrated representations of indoor, outdoor, underground and over-ground environments. While interoperability can be defined as the ability to exchange or transfer data between different applications, platforms, domains or (more generally) networks of heterogeneous systems (El-Mekawy, 2010; Eastman *et al.*, 2011; Kensek, 2014), integration may be defined as the combining of data from different sources into a single, unified view (omni.sci, 2020; talend, 2020) or environment. Much of this integration takes place at the intersection of the indoor and outdoor environments, which is approximately where the geospatial and BIM domains often collide but sometimes merge or blend (Figure 2.3.3-1).

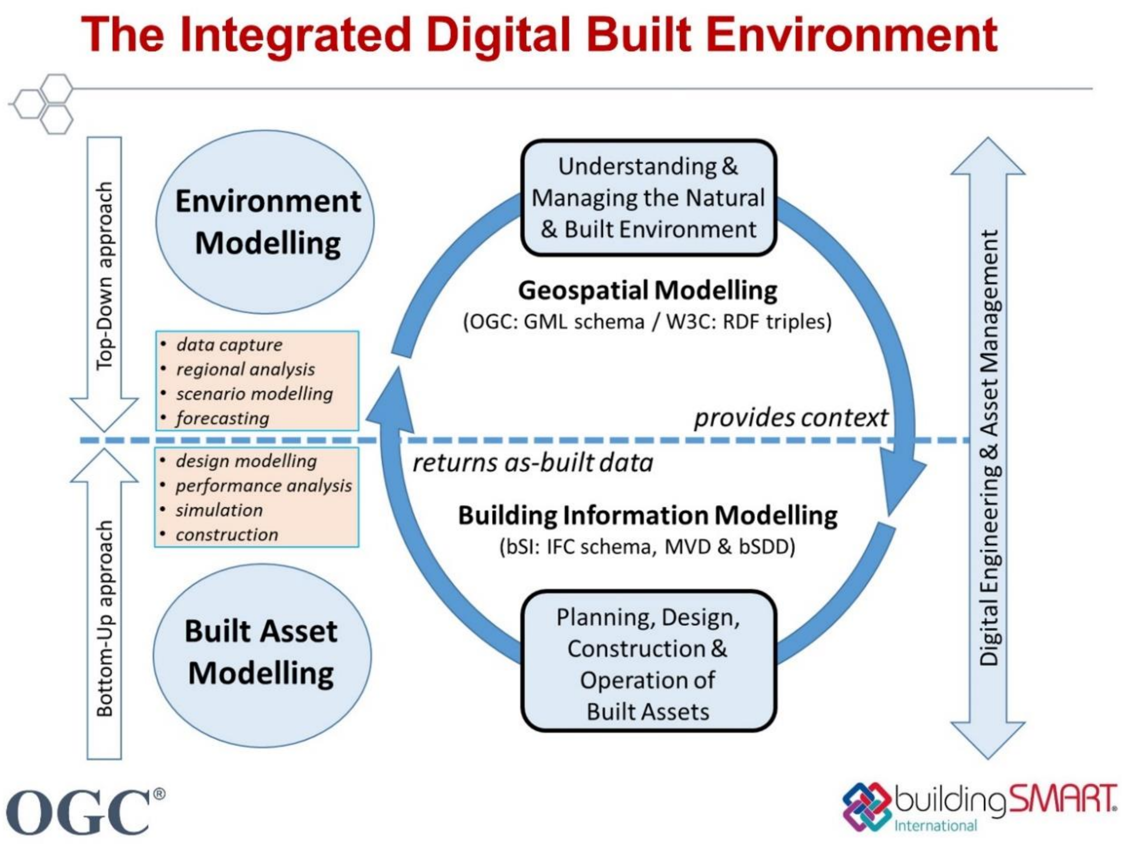

Figure 2.3.3-1 - The IDBE working group's representation of the interplay of the geospatial and BIM domains. Image credit: Jim Plume, IDBE co-chair and director at buildingSMART Australasia.

The professional practices in which urban GIS has prevailed have remained largely distinct from those for BIM, leading to some differences in the objects and environments that each is best suited to representing. BIM is primarily 'prescriptive' (meaning that it prescribes what will be built rather than describing what has been built, which would be 'descriptive') and high-detail, with datasets developed manually through a design process, optimised for modelling new constructions over relative small spatial extents using local engineering coordinate reference systems (CRSs) that largely disregard terrestrial curvature (Zhang *et al.*, 2009; Rafiee *et al.*, 2014; Tobiáš, 2015). These qualities enable BIM technology to satisfy the high precision, geometrically detailed and semantically rich modelling requirements of the AEC domain. Conversely, GIS is traditionally more descriptive, intended for modelling existing objects or environments (through automated or semi-automated processes such as photogrammetry, laser scanning and transformations from 2D landscape models) with sparse or incomplete information, often for objects and spaces which are larger than those used in BIM. GIS provides extensive spatial analysis functionality and uses absolute, geographic CRSs and map projections (Wilson and Fotheringham, 2008; Longley *et al.*, 2011; Tobiáš, 2015). These qualities enable the use of GIS technology to abstract, generalise or aggregate data that would otherwise be too large and complex to be managed and processed effectively; they also enable GIS practitioners to conduct the geospatial analyses that generally operate at a coarser scale than those of BIM.

Because of these dissimilar backgrounds and purposes, the domains have evolved divergently and produced data standards that are often disparate and incongruous. The importance of this is becoming more apparent as the need to represent environments at the boundary of the two domains – where they contribute complementary information – becomes more pressing (the use cases of section 2.2 are examples). The domains are functional and effective within themselves but sometimes restrictive when there is a need to integrate their data. An ability to address many of the use cases for multi-scale digital representations depends on the ability to integrate distinct representations of different parts of the built environment but geospatial and BIM data are underpinned by data standards that exhibit significant differences: conceptualisation of real-world objects, identification and attribution of these objects, and the techniques used for representing both spatial topologies and network topologies all differ, which makes the task of integration complicated, time-consuming and error-prone (Wang, Pan and Luo, 2019; Gilbert *et al.*, 2020; Herle *et al.*, 2020). These dissimilarities can be understood better by analysing a representative subset of the standards that are of relevance to urban data integration – the following sections presents such an analysis. Section 2.4 then interprets these disparities in context of methods that need to overcome them in order to integrate the topology of utility networks.

2.3.4 Prominent data standards at the GIS-BIM interface

Within the two domains, a broad range[15] of complex data standards[16] for the built environment has been developed. These standards vary not only in the details of their design but also their broader purpose: examples include guidance of best practice such as ISO19650 and CDB

[15] Exhaustive descriptions of these standards can be found in (National Institute of Building Sciences, 2007; Eastman *et al.*, 2011; NBS, 2011; Barnes and Davies, 2014; British Standard Institution, 2014; Kensek, 2014; buildingSMART alliance, 2015; OmniClass, 2017; WBDG, 2017).

[16] A data standard is a documented agreement on the representation of data (EPA, 2019) and is often formalised by a data schema, which defines a structure for the storage or exchange of information. An instance model then uses an encoding to store the data in a way that is consistent with the rules of the schema. Instance models are thus subject to the constraints of both the encoding language and the schema (Eastman *et al.*, 2011), allowing software that is familiar with these to understand and make use of instance models.

Spatial and Coordinate Reference Systems Guidance (Reed and OGC, 2018; BSI Group, 2021), specification of metadata requirements such as the INSPIRE standards and the National BIM library objects (European Commission, 2013; NBS, 2021) and detailed conceptual modelling with a technical schema such as GML, CityGML, IFC and LandInfra (OGC, 2012, 2016; buildingSMART, 2021). It is necessary firstly to scope standards of relevance to this thesis. ISO19650 is a recent and important series of international BIM standards that defines good practice throughout project and asset lifecycles, superseding some British Standards and Publicly Available Specifications (PAS). Although ISO19650 "considers all information whether it's a construction programme, a record of a meeting, a geometrical model or a contract administration certificate" (UK BIM Alliance, 2019), it cannot be instantiated to produce a representation of the built environment; it is not accompanied by a schema against which an instance model of a building or network asset, as examples, can be validated. The same is true for the COBIE specification, which "denotes how information may be captured during design and construction and provided to facility operators" (East, 2007), and the geospatial ISO19115 metadata standard, which "defines how to describe geographical information and associated services, including contents, spatial-temporal purchases, data quality, access and rights to use" (GIS Standards.EU, 2018). Despite the importance of these standards and that many of them underpin relevant others, this chapter is concerned with the standards that have corresponding schemas (defined data structures) that may be instantiated as machine-readable datasets. It is through the instantiation of such schemas that utility network components are represented digitally, and these digital instances are the subjects of integration.

Commonalities between standards that have instantiable schemas are opportunities for integration of their instances and dissimilarities present an obstruction: hypothetically, if two instance models are valid against the same schema, their integration may be a merge operation on the objects represented. Conversely, if two instance models are represented by standards that differ in every respect and to the fullest extent, there may be no means by which they can be compared – they would be entirely incongruous and could not be represented in a single environment. Some level of harmony between standards is thus necessary for integration. An objective is to analyse relevant standards and identify the commonalities that can be exploited and the disparities that existing methods of integration need to overcome. However, those standards of relevance are numerous and complex.

Selection of a suitably representative subset of the in-scope standards as proxies for the standards more generally is both meaningful and more manageable than an attempt to consider all standards of relevance. Two noteworthy candidates are LandInfra and IndoorGML. Introduced as a successor to LandXML, LandInfra addresses a capability gap in the modelling of land and engineering infrastructure facilities (OGC, 2016; Kumar *et al.*, 2019). IndoorGML is an extension of Geography Markup Language (GML) that serves indoor navigation applications, defining information such as spatial subdivisions, types of spatial connectivity and logical navigation networks (OGC, 2014, 2020). However, LandInfra is too young for an evaluation of its uptake and IndoorGML is too narrowly focussed on interior spatial information. Critically, neither is intended for the representation of utility networks at the scale of the building envelope. The CityGML and IFC standards, having been conceived for detailed city modelling and the communication of BIM construction designs, respectively, both overlap the building envelope in their remits and enable the representation of utility networks; for CityGML, this is through the UtilityNetwork Application Domain Extension (ADE) and for IFC through its domain-specific sub-schemas. They are relatively mature and extensively implemented, having already gained a strong foothold in urban data modelling communities, emerging as the predominant standards used in research on GIS-BIM data integration (El-Mekawy, A Östman and Hijazi, 2012; Stouffs, Tauscher and Biljecki, 2018; Noardo *et al.*, 2020). Beyond their partially overlapping remits, an important factor in this predominance has been that they are complementary (Sani and Rahman, 2018): in combination, the IFC and CityGML conceptual models are thoroughly representative of the various features of the built environment. The two standards have been analysed and compared in numerous studies on the integration of urban data for various purposes (e.g. Hijazi *et al.*, 2009; El-Mekawy, 2010; El-Mekawy, Ostman and Shahzad, 2011; El-Mekawy, A Östman and Hijazi, 2012; Cheng, Deng and Anumba, 2015; Deng, Cheng and Anumba, 2016; Donkers *et al.*, 2016; Kumar *et al.*, 2019; Gilbert *et al.*, 2020), and they are chosen similarly in this chapter as the representative proxies. Figure 2.3.4-1 provides an overview of the thematic scope of each – the type of objects and environments that each standard can represent. The figure shows how the two standards overlap in their ability to represent objects and environment across multiple themes and scales – from the finer scale of internal building furniture to the coarser scale of terrain features – but differ significantly in their strength or suitability for representing these themes or scales. At the scale of the building envelope, they may be considered equally strong or suitable.

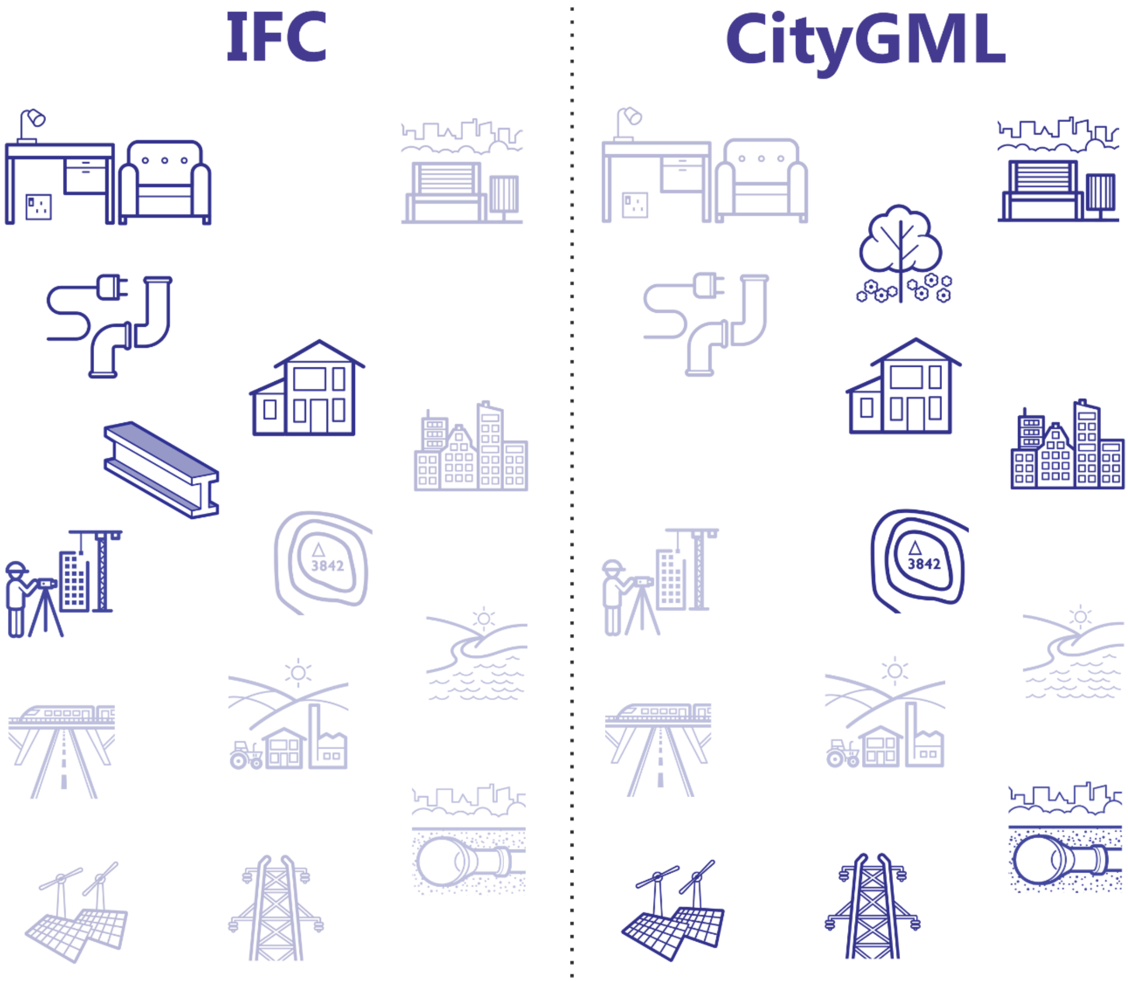

Figure 2.3.4-1 – The types of object and environment that each of IFC and CityGML are capable of representing; bolder icons indicate better representation within the respective standard. The image is an edited version of one from the Integrated Digital Built Environment working group's discussion paper (Gilbert *et al.*, 2020). The meaning of the icons is open to interpretation but the infographic is intended to convey (among other concepts) that both standards can represent buildings and utilities. IFC is stronger at the internal scale and CityGML is stronger on the city or urban scale but they overlap thematically and spatially, with both able to represent features inside and outside buildings/facilities. They are equally strong at representing objects at the scale of the building envelope.

IFC is an open BIM[17] data standard developed and maintained by buildingSMART International (bSI) and is used primarily for the representation of fine scale objects such as the detailed parts of buildings. It was conceived in 1994 to support the exchange of detailed building models (Laakso and Kiviniemi, 2012) that were developed in the proprietary formats that dominate AEC modelling practices. IFC enables much freedom in the means of representation

[17] This is intended as a reference to the openness of IFC rather than a reference to buildingSMART's openBIM (buildingSMART International, 2020b).

geometries, relationships, properties and other semantics about objects in the high detail required the AEC domain. CityGML is an open geospatial data standard for city-scale modelling that is overseen by the Open Geospatial Consortium (OGC). Development on CityGML began in 2003 following the introduction of 3D representations to GML (in version 3) and in response to various cities' and companies' inability to develop simulations on top of their 3D city models (Gilbert *et al.*, 2020)[18]. It is a coherent spatial-semantic[19] information model that defines classes and relations for objects in urban areas; it represents the geometric, topological and semantic aspects of city objects (Kolbe and Groger, 2003; Kolbe, Groger and Plumer, 2005). Given its conception in the GIS domain, CityGML is optimised primarily for descriptive modelling; conversely, IFC sits firmly within the BIM domain and, as such, is primarily prescriptive (Gilbert *et al.*, 2020). Furthermore, while CityGML was designed to serve as native or working schema (one in which modelling data is stored for intended simulations and analyses), IFC was intended originally for transfer or exchange of data between collaborators (Laakso and Kiviniemi, 2012), with industry software and their proprietary formats[20] used for BIM model development.

2.3.5 Disparities and commonalities between the data standards

The fundamental differences in origin and purpose of the standards have led to differences in their composition, some of which are readily surmountable but others of which significantly impede integration (British Standards Institution, 2014c). Focussing on the IFC and CityGML standards, Tobiáš (2015) identifies the main GIS-BIM integration obstacles as differences in semantics, coordinate reference systems and the parametric geometries of BIM - Constructive Solid Geometry (CSG) and Sept Solid representation. Gilbert *et al.* (2020) use a comparison between IFC, CityGML and LandInfra to categorise the differences with respect to the built environment domain as: intended general purpose, practical applications and modelled objects types; conceptualisation of real-world objects, their properties and their relationships; formal languages used for conceptual modelling and the description of schemas; and spatial representation (both geometric and geographic).

[18] This is known through discussions with originators of the CityGML standard, the key points of which are captured in this discussion paper.
[19] "…coherence in the geospatial context describes consistent relationships of spatial and semantic entities… if semantic and geometric aggregations show the same structure, they will be considered coherent… the more aggregation relations from concrete model instances [that] can be mapped from the geometry hierarchy to the semantics hierarchy (and vice-versa), the higher is the degree of coherence" (Stadler and Kolbe, 2007)
[20] An example of these are the Autodesk Revit software and its RVT format (the software allows export in IFC).

Similarity in intended purpose is a basis for seeking integrated representations. The use of formal languages or encodings is not fundamental: the language used for conceptual modelling and description of schemas is a matter of communication preference and compatibility of software with specific encoding languages rather than a fundamental integration criterion. However, conceptualisation, spatial representation and the expression of relationships require further analysis in order that the type and extent of limitations and opportunities they present to integration can be understood.

Figure 2.3.5-1 shows an example of a difference in semantics or conceptual modelling for each of the standards. The figure shows a generalised, schematic representation of a narrow selection of some common concepts relating to a building in CityGML and IFC: it shows how slabs are used in IFC to represent what is considered a ground or roof surface in CityGML; considered more broadly, "… a building in CityGML can be subdivided into semantic surfaces such as roofs, walls, doors, and windows. In IFC, it would instead be subdivided into the elements used in its construction, such as slabs, columns and beams, as well as fittings like windows, stairs and doors." (Kumar *et al.*, 2019). Another example is the representation of a 'space' within IFC, which has no direct equivalent in CityGML (a room in CityGML does not capture all possibilities of IFC space instances), and the composition of building as an aggregation[21] of parts in a spatial structure is a characteristic of IFC that is not present in CityGML (within the spatial structure of the IFC schema, parts of a building may exist without the existence of a building). The modification of CityGML to include IFC-like spaces has been a topic of discussion within the OGC CityGML Standard Working Group (SWG). While differences in conceptual decomposition is often a point of different between the two standards, there are commonalities on some conceptual levels, such as agreement on the meaning of the concept of a building object – this is also highlighted in Figure 2.3.5-1.

[21] Within hierarchies of data modelling, aggregation implies that the child objects may exist without the parent object; composition implies that the parent must exist (Visual Paradigm, 2021).

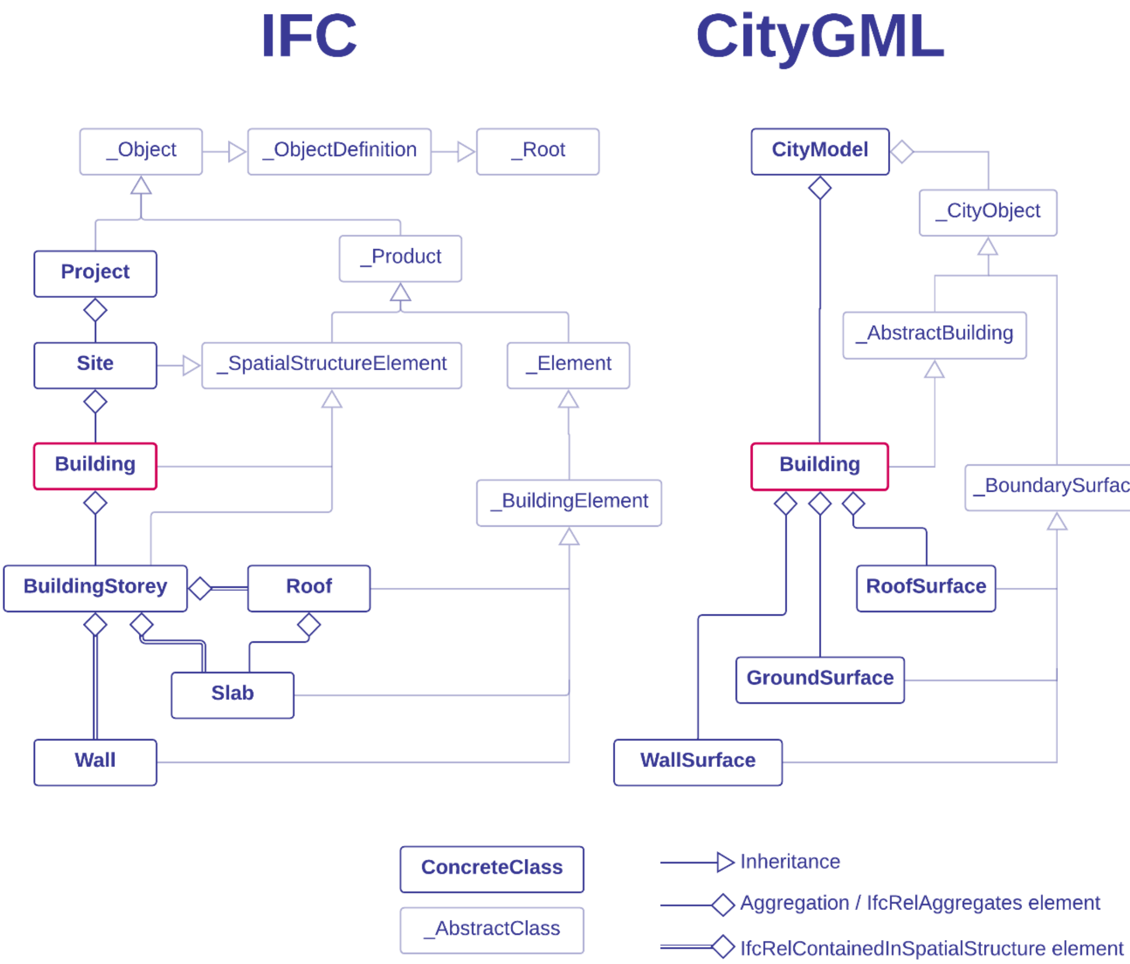

Figure 2.3.5-1 - Example difference in conceptual modelling between IFC and CityGML. The concept of a building is common but each standard represents a building's composition differently. The image is an edited version of one from the Integrated Digital Built Environment working group's discussion paper (Gilbert *et al.*, 2020). This figure shows how slabs are used in IFC to represent what is considered a ground or roof surface in CityGML.

The result of using manual design processes for the *creation* of BIM data and automated or semi-automated processes for *collection* of geospatial data is that the use of geometric representation techniques is also not always consistent between the standards. Remote sensing techniques collect data on observable surfaces but designs of future builds can model the full-depth composition of solid structures. Consequently, whereas CityGML supports (by design) only Boundary Representation (B-Rep) based on the ISO 19107 geometry model (with the restriction that only planar and linear geometry types are used), IFC geometries are based on ISO 10303 (Kumar *et al.*, 2019) and support (in addition to B-Reps) the parametric modelling techniques of Constructive Solid Geometry (CSG) and Swept Solid representation (Figure 2.3.5-2). The importance of this can be generalised by considering geometric conversions between the geospatial and BIM domains: "…enriching construction designs with real-world observation data is frustrated by the dilemma that visible object boundaries are often the only observables, which may be insufficient for the volumetric, parametric representations demanded by architects and construction engineers." (Gilbert *et al.*, 2020).

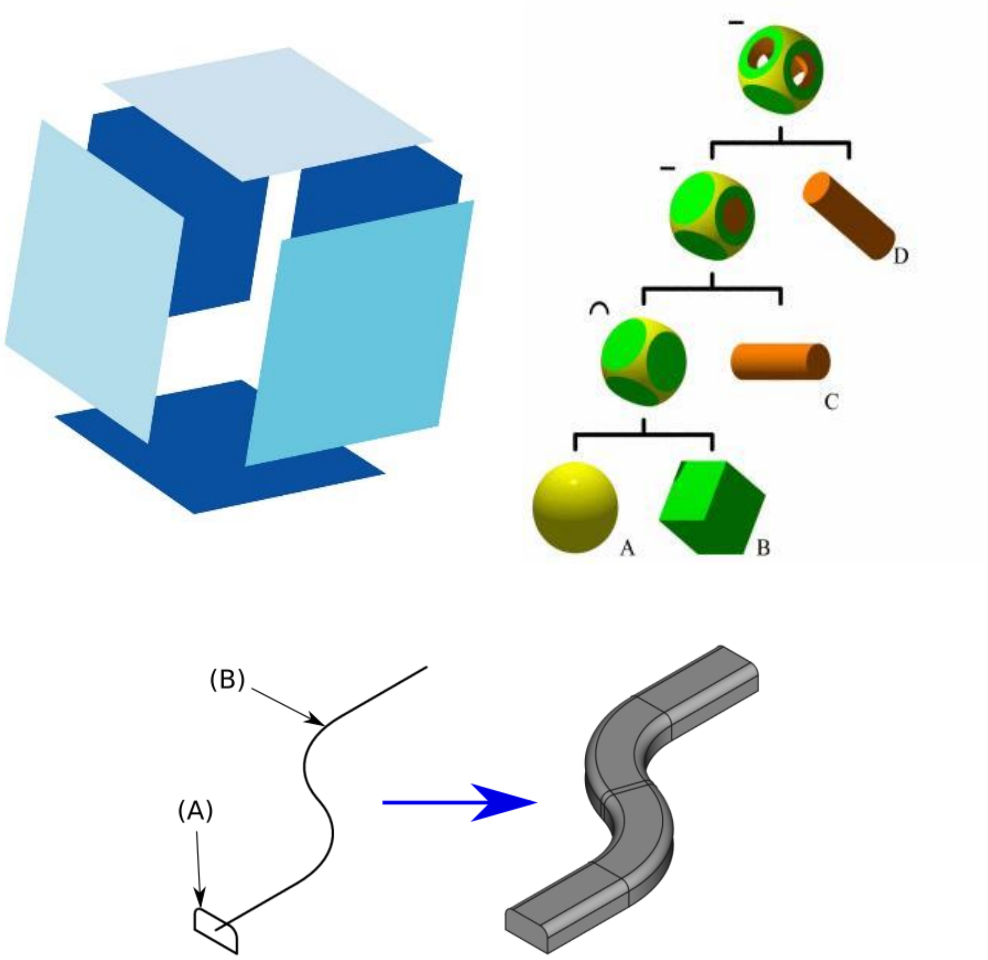

Figure 2.3.5-2 – Three different techniques for representing geometries. [top-left] Boundary Representation (Brep) (Marlow, 2005); the cube is formed by representing its boundary as a composition of six squares of different orientations and offsets. [top-right] Constructive Solid Geometry (CSG) (Goldfeather, 2015); the shape at the top of the tree is constructed by a union of sphere A and a cube B, from which two cylinders (C and D) of different orientations are then deducted. [bottom] Swept Solid (SS) (Free CAD Web, 2020); the curved pipe-like shape is created by sweeping the surface (A) along the line (B) while maintaining the surface perpendicular to and centred on the line.

In the way they account for real-world positioning, the two are also dissimilar. CityGML makes use of the geographic representations available through GML for feature representation or for anchoring features that are represented in an engineering CRS (OGC, 2012). However, given that geospatial information is not usually critical to the primary purpose (which is design) of architectural models (Diakite and Zlatanova, 2020), BIM standards generally use Cartesian engineering CRSs to represent objects. IFC enables the description of objects' geographic location through several classes (Diakite and Zlatanova, 2020) such as IfcGeometricRepresentationContext and IfcSite – Uggla and Horemuz (2018) offer a detailed analysis of these classes and the georeferencing capabilities of IFC. Again, this difference is consistent with broader patterns in the two domains (described in section 2.3.3): there is a trade-off between real-world positioning and fine resolution of spatial representation, with CityGML (and generally GIS) favouring the former but IFC (and generally BIM) the latter.

Dependent on both the location and shape of objects is topology (or spatial topology[22]), which is concerned with "qualitative properties that characterise the relative positions of spatial objects…" (Schneider and Behr, 2006). IFC defines geometries and then expresses spatial topologies through objectified relationships (such as RelContainedInSpatialStructure) between objects that use these geometries – this is related to the earlier consideration of IFC's conceptual decomposition into spatial structures. In CityGML, however, spatial topologies "…are represented implicitly in the elements' boundary representations, and composite objects are constructed using XML's hierarchical graph structure and GML's XLink mechanism. For example, a room could be represented as a composition of several polygon elements that are each defined as a linear ring (a closed loop of straight lines), grouped together as a set of XLinks references." (Gilbert *et al.*, 2020). Both CityGML and IFC exhibit geometric-semantic (or spatial-semantic) coherence (Kumar *et al.*, 2019) (the hierarchical decompositions of semantics and geometry depict the same structure) but reconciling the inconsistencies in how these topologies are expressed is a burden that integration techniques needs to overcome. The importance of spatial topological relationships for analyses of infrastructure in the built environment has been demonstrated by various studies (Jiang and Claramunt, 2004; Deleuran and Derix, 2013; Emmer, 2013; Ji *et al.*, 2017; Feng and Porter, 2020).

Deriving from communication in computing, network topology has been defined as "…the description of the arrangement of nodes… and connections in a network, often represented as a graph" (Keary, 2020) or "…the way a network is arranged, including the physical or logical description of how links and nodes are set up to relate to each other…" (DNSstuff Staff, 2019). The representation of network topology is particularly important for utility networks given the

[22] Spatial topology – also known simply as topology – is concerned with geometric properties that are preserved under continuous deformation. Spatial topological relationships include intersect, touches, contains, covers and disjoint.

complex relationships between assets and the interdependencies of different utilities (see section 2.2.4). IFC uses concrete (non-abstract) subtypes of IfcRelConnects elements to represent connections between components of a building's MEP networks; in the CityGML UtilityNetwork ADE, InterFeatureLink elements provide a similar function. This similarity represents a level of commonality between the standards. However, although 'within-model' network topologies are often well represented, 'between-model' or 'inter-dataset' topologies are not. In fact, the situation is more challenging than this: there is no standardised mechanism for expression of connections between subnetworks that are represented in instances of different standards. For example, there is no standardised means by which the network topology of a dataset representing a mains water or electricity network inside a building can be integrated with that of the exterior supply network when the datasets use different standards.

The identification of instances of concepts (the objects) within the two standards differs in scope of uniqueness: whereas IFC enforces globally unique 128-bit number identifiers for all object instances, CityGML's dependency on GML (and hence XML) requires only that an identifier begins with a letter or underscore and is unique within the scope of the instance document but not necessarily globally/universally (Portele *et al.*, 2007). However, even in cases where globally unique identifiers (GUIDs) are used throughout all source instance models, individually unique *but still different* identifiers may be used for the same real-world object. This situation is further complicated by capabilities such as multiple representations of a single real-world object in IFC, with each one suitable for a different purpose (architecture or engineering, for example).

Many of the characteristics and features of the CityGML and IFC standards are broadly representative of those of the 3D Urban GIS and BIM domains more broadly. Conceptualisation, spatial representation and the expression of relationships are information types that are key to the integrability of utility networks. Factors that are likely to hinder this integration are differences in decomposition of concepts, contrasting use of georeferenced engineering CRSs and geospatial CRSs for real-world positioning, greater variety and complexity of geometries in BIM, differences in the structuring of spatial topology and the inconsistent format and scope of uniqueness applied to object identification. Overlap in thematic and spatial remit, some commonalities in high-level concepts and some consistency the expression of internal-dataset network topology are likely to be critical to achieving integration.

2.3.6 Spatial data uncertainties

Spatial information is fundamental to both BIM and 3D GIS. Given that BIM is used to represent structures that have a location, datasets from both domains have a geographic component and "…uncertainty is inevitable in all geographic datasets and analyses." Quality, uncertainty and error are terms that all point towards a deviation from what is considered a truth: "Spatial data quality is defined based on the assumption that there is geographic truth to compare with a dataset – the closer a spatial dataset is to the truth, the higher its quality. The term 'error' refers to how far a measurement is from truth." (Li *et al.*, 2017). There is growing interest in the topic of spatial data uncertainties and its importance to utility assets – especially those situated underground – is increasing with the densification of urbanised areas.

The American Society for Photogrammetry and Remote Sensing (ASPRS) has published guidance on positional accuracy standards for geospatial data (ASPRS, 2014) and PAS128 described types of survey of underground utility detection and specific location accuracies for the measurements of type 'detection' and 'verification' (Institution of Civil Engineers, 2014). However, machine-readability is critical and neither offers a standardisation of *how* numerical accuracies should be expressed in a dataset. Hanus, Pęska-Siwik and Szewczyk (2018) discuss their analysis of positional errors for cadastral parcel boundaries, emphasising the importance of these error attributes for quality attribution, but do not discuss any convention or standard by which these errors are represented. Yan, Van Son and Soon (2021) describe how the location of subsurface utilities is critical for the management of subsurface spaces. The researchers assert that the management of acquisition, ownership, planning and development of land "…depend on and benefit strongly from the availability of reliable information of sufficient quality on the underground including utilities…", that "…much of [the] currently available data on underground utilities is of insufficient quality – in particular of insufficient locational accuracy.", and that "…the quality or lack thereof is often unknown or undocumented and may lead to inappropriate use of the data in planning and land administration decision making processes." The researchers also state that the quality of location accuracy, currency and completeness should be described using a classification system "...where data quality needs to be related to directly to its suitability to support certain processes."

Classification-based representations of uncertainty have been implemented. The PAS128 specification was introduced in 2014 to provide a standardised way to map underground utilities (Institution of Civil Engineers, 2014), capturing location accuracy within confidence categories or quality levels (QL); each QL is associated either with a description or numeric error bound (for example, QL-B2 indicates that the location has a confidence of ± 250 or ± 40 % of the detected depth, whichever is greater). The impact of PAS128 was investigated through a trial: multiple companies surveyed various underground utility assets, representing the results using the specification; the site was excavated and the results compared both with each other and the known location of the assets (Metje *et al.*, 2020). The study concluded that, although the impact of PAS128 has been positive, one of the classification levels needed revision. It is not clear whether the solution does necessarily lie in a classification system that is aligned to known and current use cases or, instead, an absolute, quantitative representation that can be applied more generally.

An example of a more general, standardised representation of location uncertainty is the OGC Abstract Specification, which includes the topic of referencing by coordinates (OGC, 2019), setting out 'ensemble accuracy' and 'coordinate operation accuracy' attributes. The ensemble accuracy is defined as the "inaccuracy introduced through use of this collection of reference frames or datums… It is an indication of the differences in coordinate values at all points between the various realizations that have been grouped into this datum ensemble". The coordinate operation accuracy is intended to indicate the error introduced through a coordinate transformation. However, the values assigned to the parameters to not appear to be formalised in a machine-readable way (for example, '3 m, 8 m and 5 m in X, Y and Z axes' is used as a data entry example) and the attributes do not apparently provide more generally for errors associated with individual observations – instead, there are specific to groups of points and transformations. Within the AEC and BIM domains, geometries and construction site positions are often represented at millimetric precision without expression of uncertainty and the relevant data standards, such as IFC, do not detail how this information should be expressed within their schemas.

A key point for this section is that, although spatial data uncertainties are well recognised as important and efforts have been made to represent them in a useful way, there does not appear to be any widely accepted, standardised method of representing such uncertainties even within one domain; there is clearly no accepted standard by which they can be represented across both the geospatial and BIM domains.

2.4 Integration methods

2.4.1 Introduction

Section 2.3 outlined the background to the geospatial and BIM domains, described some of the data standards used to represent the urban environment that is spanned by these domains, identified several features of IFC and CityGML as representative of built environment data standards more broadly and analysed these two standards to determine the key differences between them that hinder the integration of utility network topologies. Overcoming the difficulties presented by heterogeneity in standardisation demands methods and tools in support of integration and interoperability (Hijazi *et al.*, 2011). This section now considers existing methods and the latest research on integration of built environment data, categorising them by the techniques on which they are based. They methods are critiqued in the context of existing data standardisation – the commonalities they exploit and the differences they overcome – but also with respect to other ways in which they achieve target integration outcomes, such as modification to standards and development of new data standards.

Prior to BIM's emergence as a distinct discipline within the AEC industry, CAD models of buildings were recognised as suitable and important for integration with 3D urban GIS data (Liggett and Jepson, 1995; Holtier, Steadman and Smith, 2000; Benner *et al.*, 2005). It became clear that infrastructure engineering tasks relied upon both GIS and CAD capabilities and their semantic interoperability but also that the independent evolution of these two domains had given rise to differences in data formats, terminologies, semantics and techniques, and platforms that were difficult to reconcile or integrate (Peachavanish *et al.*, 2006; B. Akinci *et al.*, 2008). The increasing prominence of BIM over CAD through the early 2000s was accompanied by an increase in demand for the integration of BIM and GIS models for purposes such as building and construction analysis, urban planning, tourism, cadastre and homeland security (sensorsandsystems.com, 2008; Isikdag and Sisi Zlatanova, 2009; Laat and Berlo, 2011). The theme of focus for this thesis is utility networks and exiting methods of integration should be reviewed more broadly before their interpretation in this context.

2.4.2 Terminology and categorisation of methods

Various methods of integration are in practical use and the subjects of research efforts but there appears to be little consensus in the literature on some key definitions. Within this thesis, conceptual models are considered techniques or sets of conceptual and notational conventions;

examples are the Entity-Relationship (ER) model, Universal Modelling Language (UML) and Object Modelling Technique (OMT). Conceptual schemas, however, are specific data structures (often represented by diagrams) that are produced using conceptual models (Fonseca, Davis and Camara, 2003; Fonseca and Martin, 2007). The word 'ontology' (in computer and information science, rather than the philosophical study of being) has been used to mean the same as a conceptual schema but there is a distinction: data modellers commit to a set of computer representations when designing conceptual schemas but ontologies sit somewhere between observation and these information systems (Bishr and Kuhn, 2000; Fonseca, Davis and Camara, 2003), corresponding to how humans understand things. Ontologies are closer to the user's cognitive model (Fonseca, Davis and Camara, 2003), can provide the building blocks for conceptual models (Fikes and Farquhar, 1999) and, when formalised or implemented, are generally intended to "…provide a shared and common understanding of a domain that can be communicated between people and heterogeneous and distributed application systems." (Klein *et al.*, 2001). At a foundational level, upper ontologies are used to represent general concepts or objects; domain ontologies then specialise these for themes or application domains. Although conceptual schemas and ontologies are used extensively for the structuring of information *within* domains, both are also used within methods that seek to facilitate integration *between* domains.

A categorisation of methods enables them to be analysed and critiqued in a more structured way. The choice of categories used in other literature reviews are variable and appear to depend on the context or target use cases (Kang and Hong, 2015; Amirebrahimi *et al.*, 2016; Liu *et al.*, 2017; Floros, Ellul and Dimopoulou, 2018; Gilbert *et al.*, 2020). The following section analyses and critiques existing methods within categories that are based on the primary technique that is implemented by the methods: methods are grouped into those based on embedded referencing (section 2.4.3), direct mapping between existing schemas (section 2.4.4), the development of new conceptual representations (section 2.4.5) and the use of ontological abstraction (section 2.4.6). Finally, the use of software systems to implement these techniques is considered (section 2.4.7).

2.4.3 Embedded referencing

A level of integration can be achieved by providing a reference to an element in another dataset: the value of an attribute of an element in the 'active' dataset can point to an additional source that provides more information about the object. The reference could be a hyperlink to a website, for example, or the unique identifier of an element contained in another dataset. In an ESRI blog, Andrews (2020) conveys this in a web-to-web client integration context by describing how a "…user may be exploring a map served by ArcGIS Online and, upon clicking through an asset and investigating attribute data, may link directly to information in a BIM repository, such as BIM 360." ArcSDE serves as an interface between client geospatial software and database management systems (Esri, 1999), facilitating data transfer between BIM and GIS software by an application programming interface (API) (Amirebrahimi *et al.*, 2016). Ordnance Survey launched their Linked Data service in 2010 (Ordnance Survey, 2020a), with the updated 2013 version providing a data hub for "…access to all our Linked Data datasets, with integrated search to enable anyone to easily locate resources of interest." (Ordnance Survey, 2013). In 2015, AEC3 worked with Cardiff University on a demonstrator for the UK Environment Agency. They used the semantic web to link buildings models to public geographic (using geospatial coordinates) and weather data (using place names) to answer questions about degradation of river and coastal assets from frost (Nisbet, 2015).

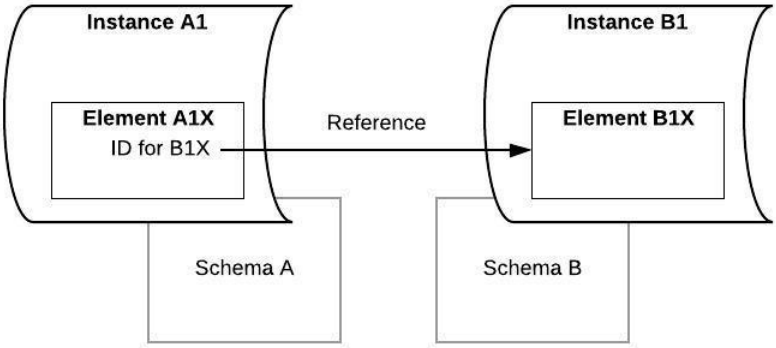

Figure 2.4.3-1 - Embedded referencing technique. A1 and B1 are instances of their respective schemas. Both instances contain elements that are representations of the same real-world object; the element A1X refers to element B1X, which contains further information on the object. This referencing allows access to information on the real-world object that is not available in A1. No schema mapping or abstraction of concepts is implied by this figure.

This embedded referencing technique provides a relatively shallow level of integration, with more depth dependent on the capabilities of any software that is tasked with making sense of the data retrieved from the referenced source but without any guarantee of compatibility or interoperability. There is also the potential for referential integrity problems – deletions won't necessarily propagate and links can 'break' (point to where the resource is no longer stored).

The techniques described in the following sections achieve a greater depth of integration by addressing differences in representation.

2.4.4 Direct mapping between existing schemas

One means of addressing such differences is to insist on all source data adhering strictly to the same set of rules or constraints outlined in one schema. By 'mapping' concepts from other schemas to the chosen working schema, all datasets can be converted to instances of this working schema. They are then valid against the single schema and thus considered integrated (or ready for integration) (see Figure 2.4.4-1). If the validity of multiple instances against the same schema is not sufficient for them to be considered integrated, this schematic consistency at least facilitates integration.

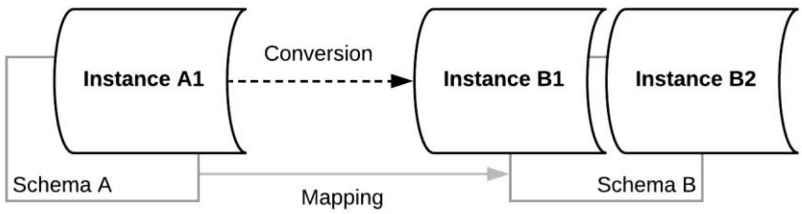

Figure 2.4.4-1 - Direct schema mapping technique for integration. Given an appropriate mapping between Schemas A and B, an instance of Schema A (A1) can be converted to an instance (B1) that is valid against a Schema B, which is then considered to be integrated with any other instance of Schema B (such as B2).

Research on custom schema mappings has tended to focus on geometries and semantics with emphasis on completeness and minimum information loss, such as the accurate mapping between B-Rep and CSG representations (Wu and Hsieh, 2007) or bringing IFC geometry into ESRI Shapefiles (Zhu *et al.*, 2019). Semantics have posed as much of a problem and Isikdag and Sisi Zlatanova (2009) presented ideas for defining a semantic mapping to allow automatic transformations between IFC and CityGML. Rizal, Michel and Pim Van Den (2013) achieved IFC to CityGML translation using CityGML's ADE mechanism but the researchers stated that the software architecture required user intervention in a number of conversion steps, questioning its efficiency and effectiveness. Cheng, Deng and Anumba (2015) addressed this efficiency concern by developing a semi-automated method that makes use of linguistic and text mining techniques to support the schema mapping process. The researchers recognise that the study assumed domain expertise and "correct and appropriate" description of entities, and that cross-domain term ambiguity could impede integration. Zhao, Liu and Mbachu (2019) demonstrate a method that involves the mapping of data into the CityGML UtilityNetwork ADE to help address space obstruction and pipe layout challenges during water network planning. However, the method does not integrate the network topologies of the source data, instead focussing on visualisation and clash detection.

Despite some momentum of effort towards completeness of integration and lossless conversions, researchers have questioned whether retention of such information richness through an integration process is necessary or even desirable. Hijazi *et al.* (2009) used a water utility network example for transforming IFC to CityGML models, stating that the "..purpose of incorporating utility networks in GIS is for modelling them topologically rather than geometrically.". Donkers *et al.* (2016) asserted that existing methods for mapping from IFC to CityGML result in models of impractical complexity by converting all geometries. They designed and implemented an algorithm for general LoD3 models that filters and maps only the necessary semantics, transforms the required geometries of the building envelope and refines the geometries to ensure model validity in CityGML. Aside from some problems with geometric conversion, the authors suggest that their mapping of semantics may not be fit for all purposes, and that an extension to the mapping depends on extension to the semantics in the IFC standard. An alternative mapping method was proposed by Stouffs, Tauscher and Biljecki (2018) for the Virtual Singapore project; they developed a CityGML ADE to extend the semantic representations of CityGML and show the use of a Triple Graph Grammar (TGG) (Schürr, 1995) for mapping from IFC into this ADE. The researchers used a set of grammar rules to generate a triad of graphs; two that represent the structures of each standard and a third that formalises their correspondences, providing a more flexible and extensible means of schema-mapping. Recognising that a fully lossless conversion may never be achieved, they state that their method allows an assessment of the completeness of any conversion and incremental development of the grammar rules that formalise the mapping. They also conclude that, although it is possible to semi-automate part of the process of mapping between IFC and CityGML, manual intervention is required for guaranteed accuracy.

An important factor in the research community's evaluation of schema mapping methods is information loss and reversibility of conversion; there is significant value attributed to less information loss and bidirectionality. Complete reversibility of conversion depends on complete information retention. An observation that can be derived from existing research is that, although schema mapping for integrating urban data can be effective and fast to develop for specific requirements, they are often then limited to niche tasks. Such a mapping technique essentially accepts and then uses existing data standards in their current form, falling short where the standards themselves prove deficient. A mapping might not be possible due to the absence of a suitably similar or equivalent concept in the target schema and, where it is possible, assumptions about equivalences need to be made. When such incompleteness or approximation of integration is considered a significant deficiency, other researchers and developers have sought to influence the underpinning representations – these contributions are now reviewed in the following sections.

2.4.5 Development of new conceptual representations

Where existing standards and the extent to which they can be reasonably or practicably integrated is inadequate, integration can be enabled through the development of a new schema or through extending existing ones. The purpose of a superset schema is to represent all of the information from subset schemas (see Figure 2.4.5-1). The superset schema can then be used as an intermediary for conversion between the subset schemas (through an indirect schema mapping) or as a 'working' representation (in place of the subset schemas) for operations such as simulations and analyses. Similarly, schemas can be extended to incorporate new features; the primary difference is that the extension inherently retains the structure and content of the original. Rather than eliminating the need for mapping into or via this new schema, the emphasis of this technique is the use of an additional or alternative structure that captures the required concepts instead of forcing a 'best fit' equivalence of dissimilar representations.

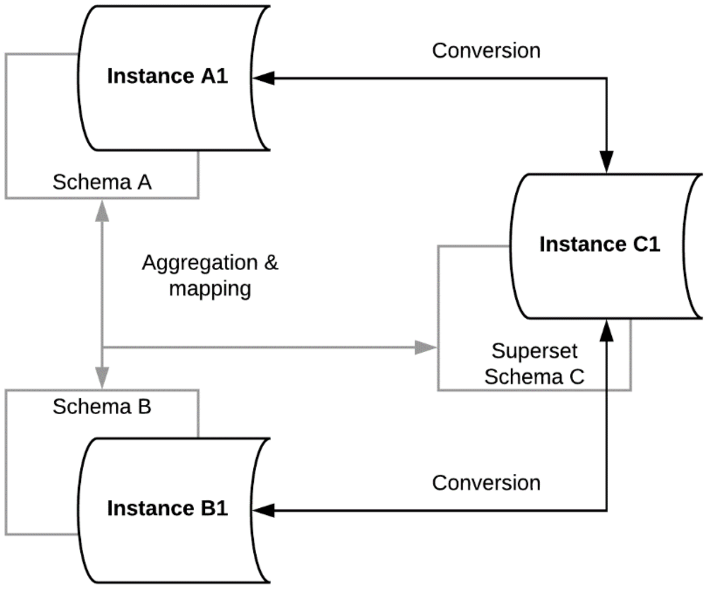

Figure 2.4.5-1 – The technique of achieving integration by developing superset schemas. Approaches that use superset schemas for integration. The concepts, geometries and other representations from schemas A and B are aggregated into superset schema C. If this process involves no abstraction, the schemas can be mapped bidirectionally and conversions can take place in either direction.

The Unified Building Model (UBM) was proposed by El-Mekawy, Anders Östman and Hijazi (2012) as a superset model designed to capture the elements of both CityGML and IFC. The researchers later focussed on utility networks (Hijazi, 2013) through development of the Network for Interior Building Utilities (NIBU) model, designed to encapsulate and provide geospatial context to utility network representations in IFC for integration with city models. However, the researchers recognise that the modelling does not include exterior utility networks and, despite demonstrating the feasibility of their proposed approaches, no evidence has been found of uptake within industry – this could be attributed to factors such as insufficient international design collaboration and consultation with stakeholders, the abrupt change that would be required from practitioners who are familiar with established data standards or challenges related to communication and promotion of new developments.

Some development of new standards has involved broader collaboration with a wider range of stakeholders. The European Commission's INPIRE Directive is aimed at facilitating international data sharing through standardisation across 34 spatial data themes, which encompass urban modelling. The schemas were extended in 2012 to include utility networks (European Commission, 2012; Vishnu and Saran, 2018). In 2020, the OGC's Model for

Underground Data Definition and Integration (MUDDI) Standards Working Group (SWG) was formed under OGC (Lieberman, 2017, 2018; Lieberman *et al.*, 2019) following concept development from three international sponsoring organisations[23] (Lieberman, 2017). The MUDDI standard derives requirements from other standards such as INSPIRE and IMKL[24], and solicits guidance from organisations that collectively have stakes in a range of use cases[25]. MUDDI includes network features and network-specific interfaces that may be used in support of utility network modelling; the IGraph interface, for example, adds relationships such as one network being a subnetwork (containment) or subordinate (dependency) of another (Lieberman *et al.*, 2019). Development of these standards has been incremental and hence slower, but their traction appears to be stronger (in the case of MUDDI, inception was too recent for any measurable uptake).

A similar effect is evident for some schema extensions and, again, CityGML is the subject of many examples; its development has enjoyed much interest from a broad range of user groups, which is due, at least in part, to the ease with which the core schema can be extended by means of the Application Domain Extension (ADE) mechanism. An early ADE example in support of GIS-BIM integration is the GeoBIM ADE (Laat and Berlo, 2011), intended to bring the semantic detail provided by IFC into the CityGML format. The researchers identified 17 IFC classes as suitable for mapping to the GeoBM ADE but discovered that the hierarchical freedom in IFC contrasts with the static structure within CityGML. They recommended the separation of geometric and semantic information, stating that the 3D geometry is "…nothing more than one of the properties of an object." and not necessary in all use cases. They also found no way of creating a network structure semantically in CityGML, such as for utilities (Laat and Berlo, 2011). The ADE that was developed in support of the Virtual Singapore project (Stouffs, Tauscher and Biljecki, 2018) (see section 2.4.4) had similar objectives. Dozens have been developed[26] but of most relevance to utility networks is the UtilityNetwork ADE (UNADE), which allows the representation of both internal and external utility networks in CityGML. Development of the UNADE followed ongoing work to bring water utility network topology BIM data into a GIS environment (Hijazi *et al.*, 2009); it was developed by international

[23] The Fund for the City of New York - Center for Geospatial Innovation, The Singapore Land Authority and Ordnance Survey (Britain's National Mapping Agency).
[24] IMKL is a common information model for cables and pipes, based in INSPIRE (Lieberman *et al.*, 2019).
[25] The use cases: routine street excavations; planning, design and construction of large-scale projects, disaster planning and response, utility related emergency response, private and public utility operations, maintenance, repair and replacement programs; and Smart Cities, Future Cities.
[26] Biljecki, Kumar and Nagel (2018) provide a thorough review of existing extensions.

consortia in consultation with various types of stakeholder and has featured frequently in publications (Biljecki, Kumar and Nagel, 2018). Despite not being developed specifically for integration with IFC, some research has investigated its integration with IFC. The UNADE does overlap significantly with IFC utility network classes and, in most cases, one-to-one mappings can be asserted (Hijazi *et al.*, 2011) without loss of information.

Modification of existing schemas such that they represent the concepts of several others supports information retention and thus removes the requirement for direct mapping between existing, heterogeneous schemas (see section 2.4.4). However, as with mapping between existing schemas, the use of superset schemas and schema extensions still ultimately depends on an ability to assert correspondence between the new or modified classes, features or elements in the subject schemas. Bradley *et al.* (2016) note the "…sheer volume of work that must be completed and further validated to fully extend a common data format for infrastructure as a whole". Although addressing some issues of incompleteness and information loss, conceptual schemas are often not close to users' cognitive models (Fonseca, Davis and Camara, 2003) and do not provide a shared understanding of built environments that spans the heterogeneous geospatial and BIM domains. This closeness to users' cognitive models is a key consideration for integration methods that target use cases for specific user groups.

2.4.6 Ontological abstraction

Ontologies are not subject to the same constraints of conceptual schemas and, although they are applicable well beyond data integration, their closeness to users' conceptualisations has made them useful in addressing some integration challenges. The principle of a minimum viable level of information retention (or acceptable level of information loss) has importance in integration methods that depend on abstraction – the process of "…hiding or removing less critical details while preserving desirable properties." (Ponsen, Taylor and Tuyls, 2010) or "…the reduction of complexity by selecting several important elements and hiding irrelevant details." (Kamarudin, Ridgway and Ismail, 2016). Ontologies can be used to abstract information from multiple sources away from *how* they are represented at source to a level at which they have a shared meaning and can be compared or equated; essentially, ontologies can be decoupled from data models. Foundational or upper ontologies may be extended using domain ontologies in a similar way to the use of ADEs for extending a conceptual schema such as CityGML.

There is a distinction between what appears to be the 'closed' nature of conceptual schemas and the 'open' nature of ontologies, which relates to closed- and open-world assumptions: the open-world assumption (OWA) asserts that something may be true even if it is not known to be true (lack of knowledge does not imply falsity), whereas the closed world assumption (CWA) asserts that what is not known to be true must be false. While a conceptual schema does not allow for a representation outside of a rigid framework, ontologies allow freedom of expression of objects and relationships that have not yet been conceived.

Despite generally being used for a higher and less constrained level of information representation, ontologies still need to be communicated in a standardised way. Built environment ontologies (amongst others) are often formalised using semantic web technologies such as the Resource Description Framework (RDF), a general purpose language for expressing information about resources (such as documents, people, physical objects, and abstract concepts with unique identifiers) in the form of subject-predicate-object triples that comprise a graph (W3C, 2011a; W3, 2014). An RDF graph can be serialised using various syntaxes[27] and some basic ontology elements are contained in the RDF Schema (RDFS) vocabulary, which consists of the specifications of classes, sub-classes, comments, and data types. The Web Ontology Language (OWL) enhances the RDFS by allowing more complexity and richness of representation; RDF graphs constructed using OWL concepts are called OWL ontologies (W3C, 2014; Pauwels, Zhang and Lee, 2017).

Relatively early studies demonstrated the use of ontologies for facilitating interoperability between the CAD and GIS domains (e.g. Burcu Akinci *et al.*, 2008) and within the BIM domain (e.g. Beetz, van Leeuwen and de Vries, 2009). More recently, a dynamic, extensible ontology that defines semantic elements, relationships and resources for 'merging' models was shown to enable the integration of georeferences and temporal data in the BIM and GIS domains (Mignard and Nicolle, 2014). Karan and Irizarry (2015) later achieved spatial and temporal GIS-BIM integration using a set of standardised construction operation ontologies in RDF; the researchers recognised the sparsity of widely accepted ontologies for construction, which has resulted in the development of multiple independent ontologies and hindering effective information transfer. Deng, Cheng and Anumba (2016) developed a superset ontology in UML for conversion between IFC and CityGML geometries but the research is limited to geometries

[27] Including RDF/XML (.RDF), N-Triples (.N_T), Turtle (.TTL) (W3C, 2011b), and Notation3(.N3) (W3C, 2011a)

(and only those in B-Rep form). Furthering their work from 2015 (Karan and Irizarry, 2015), Karan, Irizarry and Haymaker (2016) use an RDF bipartite graph model[28] to build an ontology that contains all of the relevant classes and their properties from both GIS and BIM domains. Through a case study, they demonstrate an approximate fourfold increase in feature recall (over state-of-the-art tools) during two-way data exchanges between GIS and BIM software. The authors note that multidisciplinary nature of the AEC domain leads to differences in the granularity of ontologies such that entity equivalence cannot always be asserted; they suggest identifying sufficiently similar entities could be a solution.

There is less breadth of research on the use of ontologies for utility network integration. Sánchez-De-Rivera *et al.* (2017) review existing ontologies for water management and come to a similar conclusion to that of Karan, Irizarry and Haymaker (2016), highlighting the need for a standard vocabulary. Cuenca, Larrinaga and Curry (2017) set ontology requirements for energy management and select other ontologies for merging via a core ontology; one of the merged ontologies (Gillani, Laforest and Picard, 2014) accounts for connectivity between sources and consumers but not at the granularity of physical assets and their relationships. Escobar *et al.* (2020) develop an OWL ontology to represent water supply network zones and their indicators (attributes provided by the water company) but, similarly, the ontology does not consider the connectivity of pipe assets. Although covering higher-level semantics, none of these studies addresses the integration of topological representations of utility network assets for which disparities in geometries and CRSs present an integration challenge.

2.4.7 The use of software systems and services

Although the techniques covered in the previous sections are implemented using software, the integrated outputs are generally external to the software; the software is used to homogenise source data by an abstraction and/or conversion process between standardised schemas and ontologies. Some methods achieve integration using representations internal to software, which hinges on choices made by the developers of the software rather choices made by developers of data standards. A process that involves an internal software representation can use classes of programmable objects as intermediaries for conversion between classes of open standards. For example, Autodesk Revit can export a project (an RVT file) in several formats, including IFC, DAE (COLLADA) or SKP. The SKP format is readable natively in Trimble SketchUp, which

[28] Developed by Hayes and Gutierrez (2004).

can also import DAE files. Through these export and read/import operations, the software can convert between one schema and another – in the examples above, between RVT and one of IFC, DAE and SKP. For this to be affected, schema mappings will have been set up between each of the origin and targets formats and the intermediate software representations. Software can also achieve integration through a federated approach in which the software temporarily affects a level of integration of disparate source data – Gilbert *et al.* (2020) refer to this an integration paradigm. For example, ESRI ArcGIS can import a DAE file (or an IFC file using its Interoperability Extension) and allow them to be visualised in the same geospatial environment; similarly, although not an example of integration between the GIS and BIM domains, Autodesk Navisworks allows NWD and DWF files to be combined into a single view (Autodesk, 2020), which constitutes at least a geometric integration.

The use of such software has been the subject of research on integration methods. Dore and Murphy (2012) described a design framework for integrating Historic Building Information Modelling (HBIM) that makes use of a plugin for SketchUp to bring geometric information from IFC data into CityGML. Irizarry and Karan (2012) and Irizarry, Karan and Jalaei (2013) conducted studies on optimising the location of tower cranes on construction sites and monitoring construction supply management, showing case studies that makes use of an Autodesk Revit and subsequent exports to ArcGIS. Niu, Pan and Zhao (2015) developed a web-based building energy visualisation system that uses the Revit, AutoCAD, Google Earth and EnergyPlus software in a process that converts between the gbXML, IDF, COLLADA, DXF and KML formats. Wu and Zhang (2016) show a method of importing IFC models into ArcGIS and then generating indoor route topologies from the BIM geometries.

Although often convenient for generic tasks, the use of commercial off-the-shelf (COTS) software has limitations: the effectiveness of the conversion is subject to the capability of the origin and target software applications – as are flexibility and extensibility – and any scaling up to larger datasets may incur significant manual effort (subject to the importing and/or exporting software capabilities). Although some software products allow some control over mappings (such as from RVT to IFC in Revit), niche schema mapping requirements that are critical to a specific use case often cannot be controlled without modification of the software. Given that most mappings are unidirectional with conversions incurring information loss (the importance of which depends on the use case), such limitation renders COTS ineffective for some integrations that depend on asserting equivalences that are specific to less common requirements (utility network topology integration is an example). These circumstances demand development of custom software systems or services.

Researchers have considered the merits of software system-based methods but with focus on aspects that are not clearly relevant to the goals of this thesis. Lapierre and Cote (2007) describe work undertaken at the OGC's Testbed Phase 4, which showed the feasibility of integrating CAD, BIM and GIS data for querying and visualisation via open web services; Döllner and Hagedorn (2008) describe the use of a service-based virtual 3D city model system that was developed during the same testbed. More recently, Kang and Hong (2015) use a system-based extract, transform and load (ETL) approach to build an architecture which supports information interoperability for integration in facility management (although still dependent on mappings between IFC and CityGML). The researchers remark that, despite some gains in effectiveness and performance, system-based approaches depend heavily on specialist problem-solving methods and software development time, and that it is important to present to the user only the data that are required from their perspective – again, there is suggestion of the need for appropriate information loss. It is noteworthy that these studies focus on integration at the level of visualisation and querying that is based predominantly on geographic location and geometries but do not address the network topology integration challenges that are critical from the perspective of utility infrastructure.

2.4.8 Critique of existing techniques and methods

Firstly, it is unclear from the literature what is generally understood by the term 'GIS-BIM integration'. For example, the previous section referred to schema mapping as a technique used by some methods to achieve data integration but the mapping itself might be considered integration. It is perhaps accurate to describe schema mapping as the assertion of a correspondence that enables data to be homogenised and then integrated – brought into a single view or environment. Similar analysis can be applied to the other techniques. A common (or clearly defined, at least) understanding of what is meant by integration in a particular context is necessary for framing research gaps. In this thesis, integration is intended to refer to a process carried out on datasets that allows the objects they represent to be analysed and visualised as a single system within the same digital environment (this definition is refined in section 2.5 for the specific focus of this research).

There is no evidence of methods that address the use cases of section 2.2 for existing industry datasets describing real-world utility networks. The methods reviewed over sections 2.4.3 – 2.4.7 have, in some cases, been applied to utility networks or developed for utility network-related use cases. The few examples addressing the integration of utility network topologies across the building envelope involve the modification of an existing standard (such as the UtilityNetwork ADE for CityGML) or the development of a new standard (such as NIBU and MUDDI) – these are detailed in section 2.4.5. However, none of these standards has yet been implemented extensively within industry (MUDDI is in early development at time of writing this thesis and focusses on underground environments). The UtilityNetwork ADE is a strong candidate solution but, even if its uptake does increase rapidly, it is not likely that it will gain sufficient traction within both the BIM and GIS communities soon enough that it can be relied upon as a solution to many existing integration challenges. Furthermore, no method has been found that overcomes the inter-dataset connectivity issue described in section 2.3.5. There is a general lack of method for integrating existing digital representations of real utility networks for the use cases of section 2.2.

The difference between methods that depend on representations internal to software systems and those that do not relate to permanency and openness, neither of which is technically fundamental. Given that software representations can be persisted as project files and given the recent trend towards database representations of city modelling data that have traditionally been file-serialised (Yao *et al.*, 2018), permanency is not a binary condition and is not clearly a distinguishing factor. Furthermore, openness of data standards is primarily a commercial factor and thus also not fundamental on a technical level. Although software-based methods may have to deal more often with issues such as change propagation and consistency, it can be reasoned that they exhibit more commonality than disparity with other methods: the software must still interpret the schemas of the source data (and, in some cases, still make use of conversions between these schemas), abstract information where required and then structure the concepts in an integrated way. Information is still lost in the abstractions required for integration of the in-memory representations.

The techniques and their implementing methods differ primarily in the level of abstraction (and hence the amount of information lost), how explicit this abstraction is, and the depth of integration achieved. As a hypothetical example, a direct schema mapping may convert 'cable' or 'pipe' in one schema to 'conduit' in another schema that makes no reference to resource type

– this information is lost implicitly. Alternatively, the origin schema could be extended to include a parent 'conduit' class; after generalising the asset to this conduit class, the resource type is explicitly removed but the result is the same. A lossless conversion may be achieved by extending the target schema to represent 'cable' and 'pipe' along with an attribute for resource type. Embedded referencing, however, makes no attempt to reconcile differences between representations, leaving this task to the user software. For example, building footprints could serve as placeholders for further building detail, with links to a URI of the BIM data embedded in attributes of the footprint element, but this link alone does not provide a means of making sense of the building detail. For this reason, although neither the example above of schema extension nor that of embedded referencing involves information loss, the former achieves more depth of integration. The use of a common ontology for integration involves explicit abstraction to overcome an inability to assert equivalence between existing representations[29]. By simplifying complex data structures (consider the compositions shown in the UML-like diagrams of Figure 2.3.5-1) to a more intuitive set objects and relationships that are closer to the user's cognitive model (see sections 2.4.5 and 2.4.6), the use of ontologies for integration can avoid much of the complexity incurred through mapping between disparate conceptual schemas.

There has been some emphasis in the literature on the value of moving from the more prevalent partial concept matching and unidirectional conversions towards lossless methods of integration and bidirectional mappings (El-Mekawy, 2010; El-Mekawy, A Östman and Hijazi, 2012), and it has been shown that this can be enabled through, in particular, schema extensions and superset schemas. Despite this, there is now a growing appreciation of the need "…to develop integrated digital representations that enable the execution of queries, analyses and visualisations that operate with sufficient fidelity across the digital environment; instead of aiming for losslessness, the objective then is to achieve a level of information recovery that is appropriate for target use cases." (Gilbert *et al.*, 2020). Rather than seeking fully lossless integration, partial integration (or integration of partial information) is being recognised as sufficient, pragmatic and sometimes superior for some use cases (Stouffs, Tauscher and Biljecki, 2018). The partialness can be thematic, such that only a portion of an instance or schema/ontology is integrated with another, or in the level of information granularity, such that detail is abstracted in favour of homogeneity. The specific requirements of a particular user,

[29] The technique of 'smushing' can be used to map from one resource in an RDF graph to another given an equivalence/similarity predicate or similarity in property values (Dataincubators, 2012).

their use case and their cognitive model (see section 2.4.6) should be considered when designing and developing an integration method. However, it has also been assessed that much existing applied integration research has been limited to use cases that are atypical or unrealistic (Fosu *et al.*, 2015). Abstraction is critical to many modelling and integration tasks but the extent of information loss and type of information lost through such abstraction should be appropriate to realistic, current and worthwhile use cases. The measure of success of any method is the usefulness of the integrated data representations at providing the functionality demanded by real-world use cases.

Another observation is that it is not always possible to achieve a desired integration outcome with the data standards as they exist; some conceptual models and their instances are too disparate to be reconciled to an extent that they can be integrated for all or any use cases. Under these circumstances, data standards should undergo development. In March 2020, the OGC, buildingSMART International (bSI) and the Integrated Digital Building Environment (IDBE) working group published a set of action points for addressing some of the challenges in built environment integration, one of which was to agree on "…a collaborative mechanism for opportunistic harmonisation of conceptual representation…" in the relevant data standards, suggesting that a level of commonality or congruence between data standards can be leveraged for integration of their instances but recognising that "…it is neither feasible nor desirable to redesign [the data standards] from scratch.". (Gilbert *et al.*, 2020). Herle *et al.* (2020) make similar observations and conclusions, emphasising the need for coordinated standardisation activities. Enhancements to open data standards should be careful, collaborative, iterative and incremental but this incurs the risk of slow progress. Some standards require a level of implementation prior to adoption; for example, the OGC requires "strong evidence of implementation" prior to adoption of a standard as a Community Standard, which represents an official position and endorsement by the consortium (Open Geospatial Consortium, 2017). Case studies, software development and demonstrators aimed at influencing the trajectory of standards development can be innovative and disruptive, increasing the pace of development and confidence in the appropriateness of enhancements, and accelerating adoption of changes to standards in the relevant community.

2.5 Research gaps

Before targeting further research, the understanding of data integration clarified at the beginning of section 2.4.8 needs to be refined such that it is applicable to the use cases of this thesis. Integration of utility network topologies is intended to mean the process of asserting connectivity between multiple disjoint digital representations of real-world sub-networks that are physically and functionally connected in the real-world, allowing them to be analysed as a single network in a digital environment. For this integration to be successful, it must be automatable and enable the analyses and simulations required by the use cases.

Firstly, research should identify what details can be abstracted (and hence information lost) through the application of any proposed method, what information needs to be retained, how these data should be structured and the technologies that can support this. Secondly, research needs to identify how existing representations inhibit the automation of this integration and, where they do, how the underpinning data standards can be modified to better facilitate utility network topology integration.

There is little evidence of integration methods that look specifically at enabling simulations of multi-scale resource flows and analyses of favourable network topology, both of which are relevant to the use cases of section 2.2. Research has focussed strongly on geometry with relatively little consideration of network topology, and there is a weakness in current capabilities to integrate data across multiple scales. The opportunistic exploitation of commonalities between the data standards should be investigated as 'hooks' (Gilbert *et al.*, 2020) for integration of utility networks across multiple spatial scales and the suitability of methods should be evaluated against factors such as effectiveness and flexibility (Liu *et al.*, 2017) and the extent to which they can be automated.

There are two spatial factors that have been underemphasised by existing research and development: the consistent and accurate terrestrial positioning of real-world objects and a generic, standardised mechanism of representing uncertainties in these positions. The lack of georeferencing in many BIM models significantly hinders their reuse in GIS environments (Ohori *et al.*, 2018) and the consideration of accurate georeferencing is often overlooked by integration methods (Uggla and Horemuz, 2018). Due to the network topology (connectivity) of geospatially constrained utility networks being so closely related to their spatial topology, and the necessary dependency of spatial topology on geometry and real-world location, the accurate geo-referencing (or geospatial positioning) of both BIM and GIS objects may be an important factor in the derivation of network topologies where functional relationships between components of disjoint GIS-BIM datasets are not explicit. The standardised representation of spatial uncertainty would enable a better scoping of the suitability and limitations of methods that depend on a minimum level of spatial accuracy.

The way in which these objects are identified within digital representations would also benefit from further research. Although most data standards provide a means of universal (or global) unique real-world object identification, there is neither an encoding system nor deduplication mechanism (intended as an aspect of data integrity prior to any integration) that is common to them all; unique identifiers vary in their format and multiple different identifiers may refer to the same real-world object.

With respect to integration of network topology, both the importance of object identification and real-world positioning of these objects are factors that need to be researched further.

2.6 Research approach

Existing research has focused on how techniques can be applied for generic, high fidelity integration but often without thorough consideration of the justification for the extent or type of integration. There is an opportunity to devise integration methods through an exploratory and flexible approach that iterates towards satisfying use case requirements, making use of focused case studies in which tangible problems need to be solved with influence from data producers and application domain experts.

It is important to repeat and further explain the difference between the intended definitions of use cases and case studies in the text of this thesis. It is also important to distinguish between what is meant by methods and techniques – again, in the context of this thesis. As explained in section 2.2.1, use cases are understood to be more generalised scenarios than case studies, which are real-world realisations of the use cases. It is also important to distinguish between what is meant by techniques and methods. While integration methods implement one or more techniques, the methods exist to satisfy the requirements of a use case, a better definition and understanding of which may be gained through case studies. For example, a method may implement schema mapping (technique) within a specific workflow that homogenises and integrates spatial data (method), which enables the visualisation of a multi-scale walk-through simulation (use case), the specific requirements of which could be understood by studying the projects of particular architects and their clients (case study). Case studies can be used to inform method design, which is a key activity within this research. This structure is depicted in Figure 2.6-1.

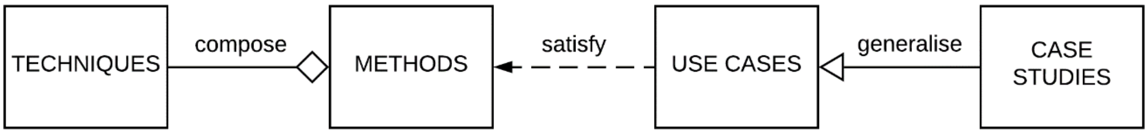

Figure 2.6-1 – A structure used for the research of this thesis. Techniques are implemented by methods, which satisfy use cases, which may be understood through case studies. The structure can be used to reason that case studies inform method design and choice of technique.

Philosophically, this research targets the research gaps identified in the previous section through an approach that is based on the ideas of pragmatism and heuristics in engineering (Goldman, 2004; Martinelli, 2012). The approach also places importance on consultation with data producers and owners, and application domain experts through the case studies and the researcher's membership of the Integrated Digital Building Environment (IDBE) working group – an international initiative under buildingSMART International (bSI) and Open Geospatial Consortium (OGC), which aims to "…achieve better software interoperability and data integration in the geospatial and built environment domains through coordination of standards development activities." and advocates "…fostering participation from a wide range of stakeholders and a common understanding of the key problems and objectives." (Gilbert *et al.*, 2020).

Technically, the approach of this project involves the exploration of case studies using iterative software development to inform the design of integration methods. Although many agile software development principles are more applicable to teams working on commercial projects, the software development in this thesis adheres to the following: a project plan must be malleable, responding to changing requirements (Amber, 2011), working software is the primary measure of progress and simplicity (the art of maximising the amount of work not done) is essential (Beck *et al.*, 2010). These principles are applied to the (software[30]) development of integration methods in support of specific case studies.

The research objectives of this thesis involve the prototyping of methods of integrating the network topologies of electricity and water networks across the building envelope at the Helix site in Newcastle upon Tyne. The Helix site is chosen opportunistically: the researcher is located on the Helix site and detailed datasets describing the new-builds and the surrounding

[30] The software developed in this research is available in the Git repositories detailed at the beginning of this thesis.

urban area are both available and accessible. The utilities of water and electricity are chosen to enable the research to address the use cases described in section 2.2 and because these utilities are represented digitally with sufficient coverage and detail.

The research questions (see section 1.4) evolved as the case studies progressed and exposed the underlying integration challenges; the use cases (see section 2.2), which serve as generalised forms of the requirements of case studies (see Figure 2.6-1), were also down-selected and refined through implementation of and reflection on the importance challenges addressed in the case studies. Throughout the case studies, persons from academia, standards bodies and utility companies were approached for discussion and guidance that supported this iteration towards appropriate research questions and use cases. These individuals are described in Chapter 3 and Chapter 4 in the context of the case studies.

2.7 Summary

Three use cases are identified for their demonstration of the benefit that can be derived from integrating network topology across the building envelope. The use case of 2.2.2 showed the value of integrated multi-scale visualisations of electricity usage to building estates managers and conscientious consumers; 2.2.3 described how real-time water network partitioning can be optimised through integration of building data with the urban scale networks; and 2.2.4 outlined how buildings and distribution networks can be better planned in the context of each other if their digital representations can be integrated.

The utility networks of these use cases span the spatial scales of the geospatial and BIM domains. Although overlapping and merging in their remits, these two domains have different cultural and practical backgrounds, which has given rise to differences in data standards that are significant enough to inhibit integration of digital representations of utility networks. An analysis of CityGML and IFC as proxies for the 3D urban GIS and BIM domains more generally is used to identify the key categories of disparity as conceptualisation, unique identification real-world objects, geometries, location, spatial topologies and between-network topologies.

Researchers and practitioners have used different techniques as bases for integration methods intended to overcome these disparities. Fundamental differences between these methods are identified as including the extent and explicitness of information abstraction and the depth of integration achieved. It is observed that complete, lossless integration of digital representations is often neither necessary nor desirable; that abstraction is critical to many modelling and integration tasks but the extent and type of information lost through such abstraction should be appropriate to realistic, current and worthwhile use cases. Where existing digital representations prove too disparate to be reconciled, data standards should undergo collaborative, international, multi-domain, iterative and incremental development; conversely, it is suggested that software development and case study demonstrators aimed at influencing the trajectory of this development can be novel, high pace and disruptive. Several technical research gaps are identified: the opportunistic exploitation of relevant conceptual commonalities between the data standards; the *common* and unique identification of real-world objects; and the accurate and consistence geo-referencing or geospatial positioning of utility infrastructure assets. The suitability of any methods designed to address these gaps should be evaluated against factors such as effectiveness, flexibility and potential for automation.

The research approach is this thesis is to address the research gaps using pragmatism and heuristics in the development of methods for case studies that are chosen to match the target use cases. The approach values consultation with data producers, data owners, and application domain experts, and active participation in relevant international forums. The technical approach embraces some agile software development principles: development should respond to changing requirements, working software is the primary measure of progress and simplicity is essential.

Chapter 3 Integration of electricity distribution and consumption networks across the building envelope

3.1 Introduction

The first use case detailed in Chapter 2 concerns electricity demand-side management (section 2.2.2). This chapter addresses this use case and the research gaps identified in section 2.5 through a case study concerning the integration of digital representations of electricity networking inside the Urban Science Building (USB) of Newcastle's Helix site and the distribution network of the surrounding urban area. The network topology of the networks is integrated, and their flows are simulated, by exploiting basic conceptual overlap between the underpinning data standards, abstraction of redundant detail and the use of graph network representations.

The method is implemented using a prototype software system that is developed iteratively while exploring the features and characteristics of the available data, making use of several prominent graph database, message broker and web technologies. The applicability of the method and technologies is demonstrated through a dynamic visualisation that supports demand-side management. The discussion considers the extent to which the data standards in their current form facilitate the method and how they can be improved where they are inhibitive.

Several stakeholders in the Helix project and (to varying extents) the outcomes of this research were consulted throughout this case study. The BIM Manager from the prime contractor (Bowmer & Kirkland) for the Urban Sciences Building (USB) contributed initial guidance on data availability and acquisition; site visits by the Senior Project Manager and engineers from NG Bailey (the engineering contractor) were used as opportunities to understand the electricity layout diagrams for the USB and data from the Building Management System (BMS). With the support of a Research Software Engineer from Newcastle University, this information was used to develop an understanding of how the utility consumption streaming data provided through the Urban Observatory’s web services map spatially to the electricity consumption within the building. The urban-scale electricity distribution network data were derived heuristically.

3.2 The Helix site

The Helix site is a £350million, 24-acre flagship project and testbed, bringing together academia, communities, business, industry and the public sector in a collaborative ecosystem (Newcastle University, 2018a, 2018c). Helix comprises multiple new-build research and innovation facilities, including the Urban Sciences Building (USB), Catalyst and Frederick Douglas Centre (FDC) (see Figure 3.2-1).

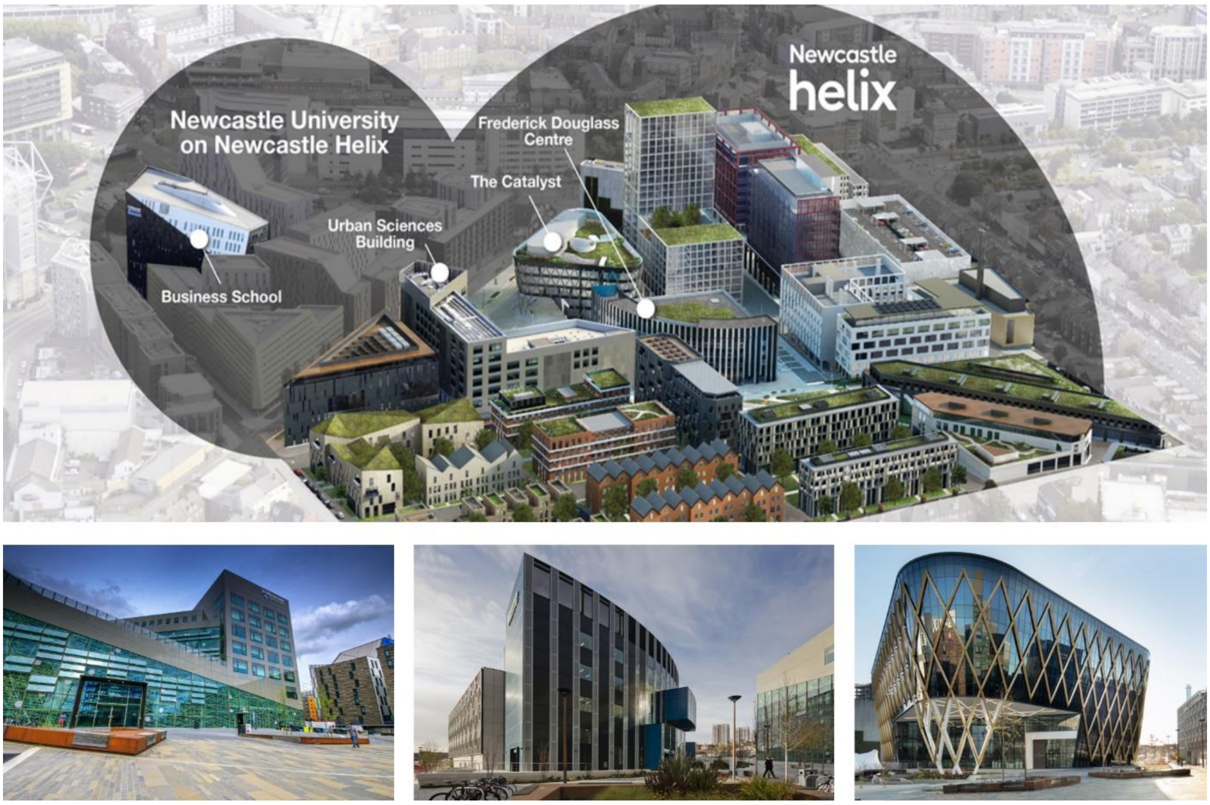

Figure 3.2-1– Top: an artistic impression of the Helix site (viewed from the Northwest) along with other existing and future constructions in the surrounding area (image credit: Newcastle University); bottom-left: Urban Sciences Building (USB, B1, image credit: Hardscape); bottom-centre: Frederick Douglass Centre (FDC, B2, image credit: Sheppard Robson); bottom-right: Catalyst (B3, image credit: Flowcon International).

The FDC is a teaching facility that includes an auditorium, lecture theatre, seminar rooms and exhibition spaces (Newcastle University, 2020). The Catalyst is home to the National Innovation Centre for Ageing (NICA) and the National Innovation Centre for Data (NICD), providing office space for businesses operating and wishing to collaborate in these centres' sectors (Newcastle University, 2021). The USB is home to Newcastle University's School of Computing and several research labs including the Centre for Energy Systems Integration (CESI) and the Urban Observatory (UO). A set of Mechanical, Electrical and Plumbing (MEP) BIM models exist for all three building, providing detailed representations of the internal utility networking. All three of these buildings are the subject of the case study of Chapter 4; this chapter focusses on the USB.

The USB is monitored in high detail by approximately 4000 sensors (Zirak, Royapoor and Gilbert, 2019), serving as "…a demonstrator for understanding the relationship between buildings and their wider environment. The thousands of sensors located in the building make it possible to not only understand its performance, but also how it interfaces with the energy, water, internet and other networks it is connected to" (Newcastle University, 2018c). These data streams are consumed, managed and then made publicly available by the UO. In combination with the BIM MEP modelling, the availability of real-time data describing its internal power consumption make the USB a suitable candidate for a study of electricity flows.

3.3 Data sources

3.3.1 Overview

Chapter 2 analysed CityGML and IFC as proxies for data standards in the GIS and BIM domains, describing some overlap and complementarity in their thematic remits. In order to align subsequent method development with the structure of a prominent data standard that is designed for the representation of utility network, it is ensured that the datasets describing the physical electricity infrastructure used in this case study are represented in CityGML and IFC – those standards used as proxies for the 3D Urban GIS and BIM domains in the analysis of section 2.3.4. Where the original source data is not represented in one of these standards (see section 3.3.2), a conversion is carried out. The intention is to maximise the applicability of the developed method to the datasets of other integration scenarios.

3.3.2 Geospatial distribution network

In the United Kingdom (UK), distribution operators manage the flow of electricity from transmission substations, through distribution networks, to end-users (UK Parliamentary Office of Science & Technology, 2014). However, there are no publicly available data that describe the spatial layout of local distribution feeder networks, which are necessary for the modelling in this case study. For this reason, a set of heuristically derived synthetic distribution networks produced by Ji *et al.* (2017) were used as a plausible alternative for the network around the Helix site. The researchers' algorithm uses Ordnance Survey (OS) Points of Interest Data for the location of electricity substation, OS MasterMap® Topography for building footprints and OS MasterMap® Integrated Transport Network (ITN) for roads to derive cable layouts between the substations and buildings (Figure 3.3.2-1).

Figure 3.3.2-1 – Heuristically derived distribution network dataset for the west side of Newcastle upon Tyne. The small blue dots indicate centroids of the buildings, which are coloured grey. The red triangles locate electricity substations. The black dots represent substation and building access nodes.

The source data for the synthetic distribution network shown in Figure 3.3.2-1 is represented as nodes and edges in the Esri Shapefile format. The nodes consist of substations, substation access nodes, building access nodes, and buildings; the edges represent cables that carry electrical current. Although alternating current (AC) electricity supply involves a high-frequency alternation of direction of electron movement, for this case study, a fixed flow direction was attributed to the graph features[31] in order to capture the demand-supply hierarchy in support of subsequent analysis and visualisation.

For reasons detailed in section 3.3.1, the distribution network data are converted to an instance of the CityGML UtilityNetwork[32] Application Domain Extension (ADE) (section 2.4.5 describes how this ADE overlaps significantly with IFC utility network classes). The conversion was achieved using a schema mapping, which was developed in a workspace of Safe Software's Feature Manipulation Engine (FME). The output CityGML network comprises NetworkGraph elements that contain multiple FeatureGraph elements, within which nodes are connected using InteriorFeatureLink elements. This structure is shown in Figure 3.3.2-2.

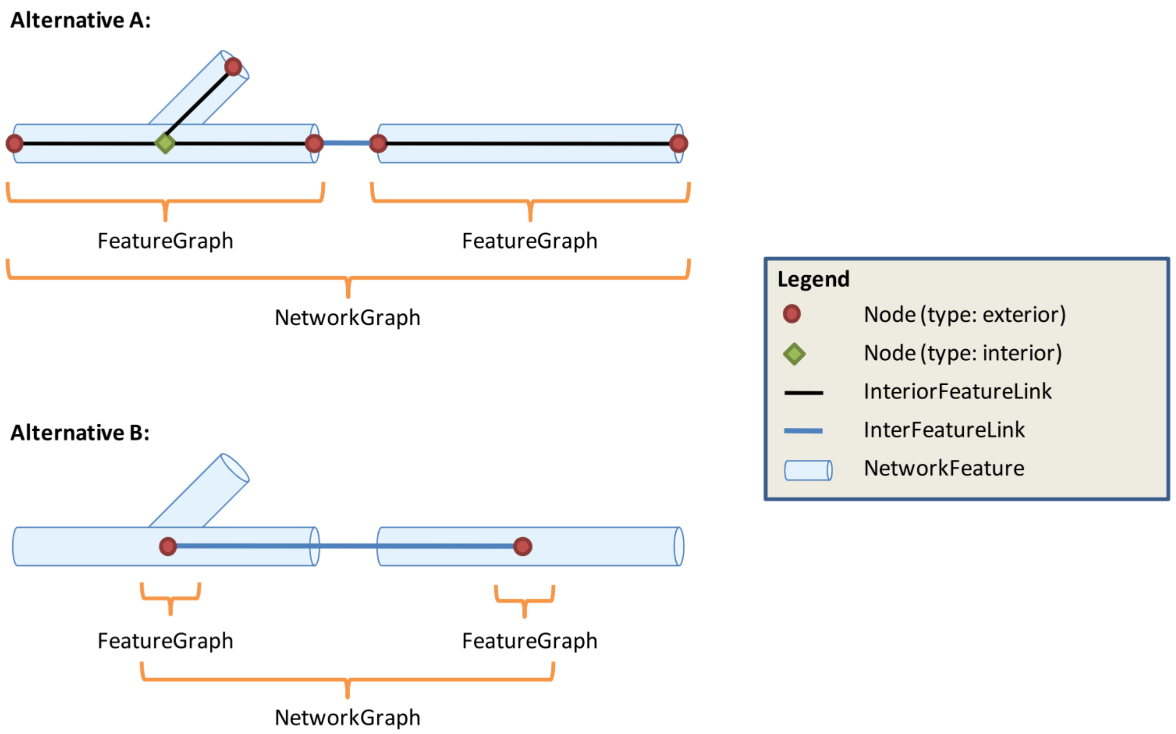

Figure 3.3.2-2 - The UtilityNetwork ADE's NetworkGraph, FeatureGraph and InteriorFeatureLink elements that can be used to represent connections between utility assets in CityGML. Alternative A is chosen as the destination respresentation for mapping of the the distribution assets from Esri Shapefiles. This figure is a portion of one presented in the core model document for the ADE (Becker, Nagel and Kolbe, 2010).

[31] This task of adding flow directions was carried out by Qingyuan Ji.
[32] More information on this ADE can be found at the ADE's wiki page (Becker, Nagel and Kolbe, 2016) and at the following references: (Kutzner and Kolbe, 2016; Biljecki, Kumar and Nagel, 2018)

The distribution network data shown in Figure 3.3.2-1 is a set of smaller networks that are supplied by substations. From this set, a single distribution network was selected for use in this study and is shown in the inset of Figure 3.3.2-3. The network comprises two building nodes, a single substation, two building access nodes and one substation access node, all of which are situation within 200 m of each other. One of the buildings represents a real building footprint from Ordnance Survey data and is labelled Building X (BX) (the real-world building represented by the footprint is not important for this study).

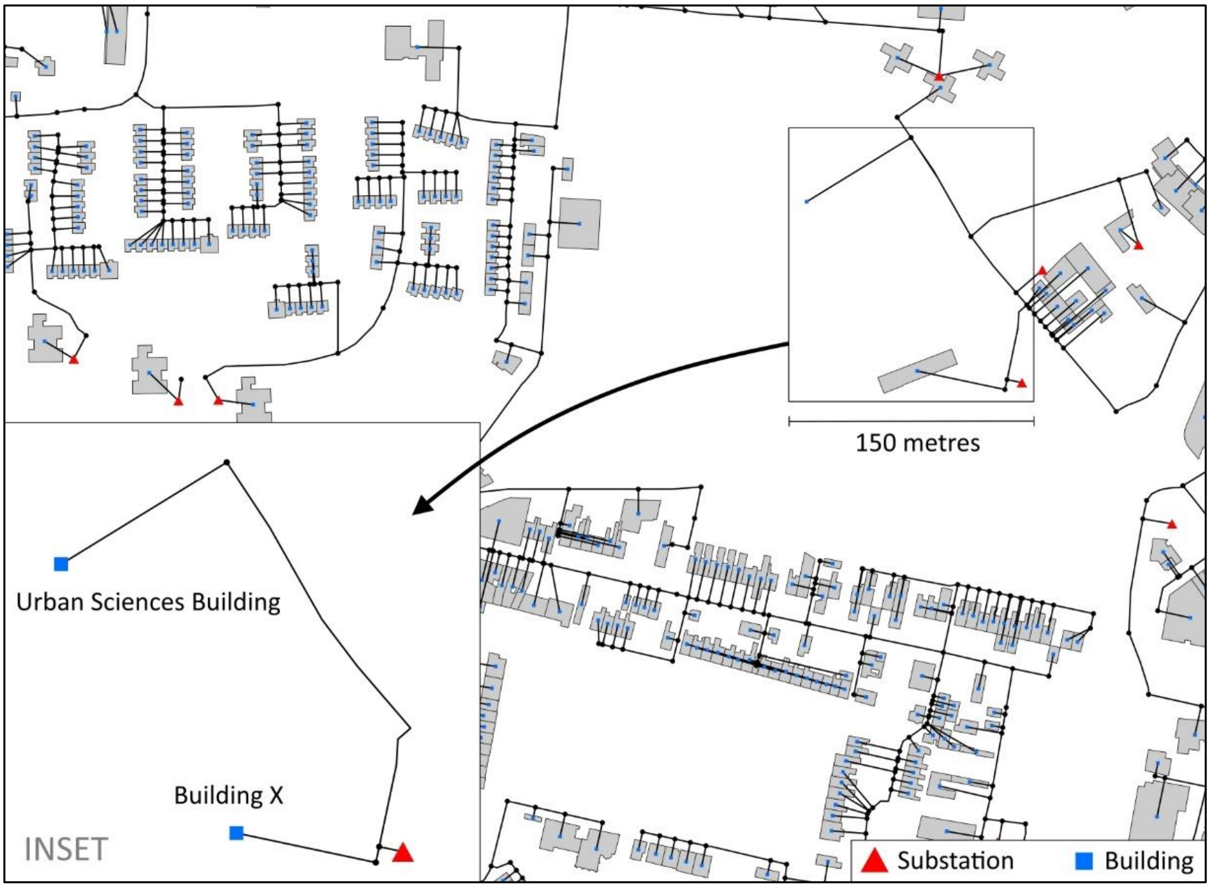

Figure 3.3.2-3 – The distribution networks closer to the Helix site: the inset region approximately encompasses the Helix site, showing the small distribution network that supplies the USB.

3.3.3 BIM MEP modelling

The two building nodes shown in Figure 3.3.2-3 are both used in this case study, each providing a different purpose. There is a need to apply any developed integration method to demonstrative, technically manageable and well understood datasets representing the static structure of a building. The BIM MEP modelling for the USB is highly complex and a simpler model is required. BX serves as a placeholder for which a simple, well understood BIM model of a building could be created. The node representing the USB serves the purpose of locating topologically and geographically the real-time sensor data streams for the building (described in section 3.3.4), providing detailed information on real-world electricity usage.

The BIM model for BX was created using Autodesk Revit 2016. The structure is a generic building comprising a single room (or space) with four walls and a pitched room. The building includes realistic objects that are familiar within both domestic and commercial settings: three floor lights, two wall-mounted screens and five mains power sockets. Figure 3.3.3-1 show the BIM with architectural features and the appliances.

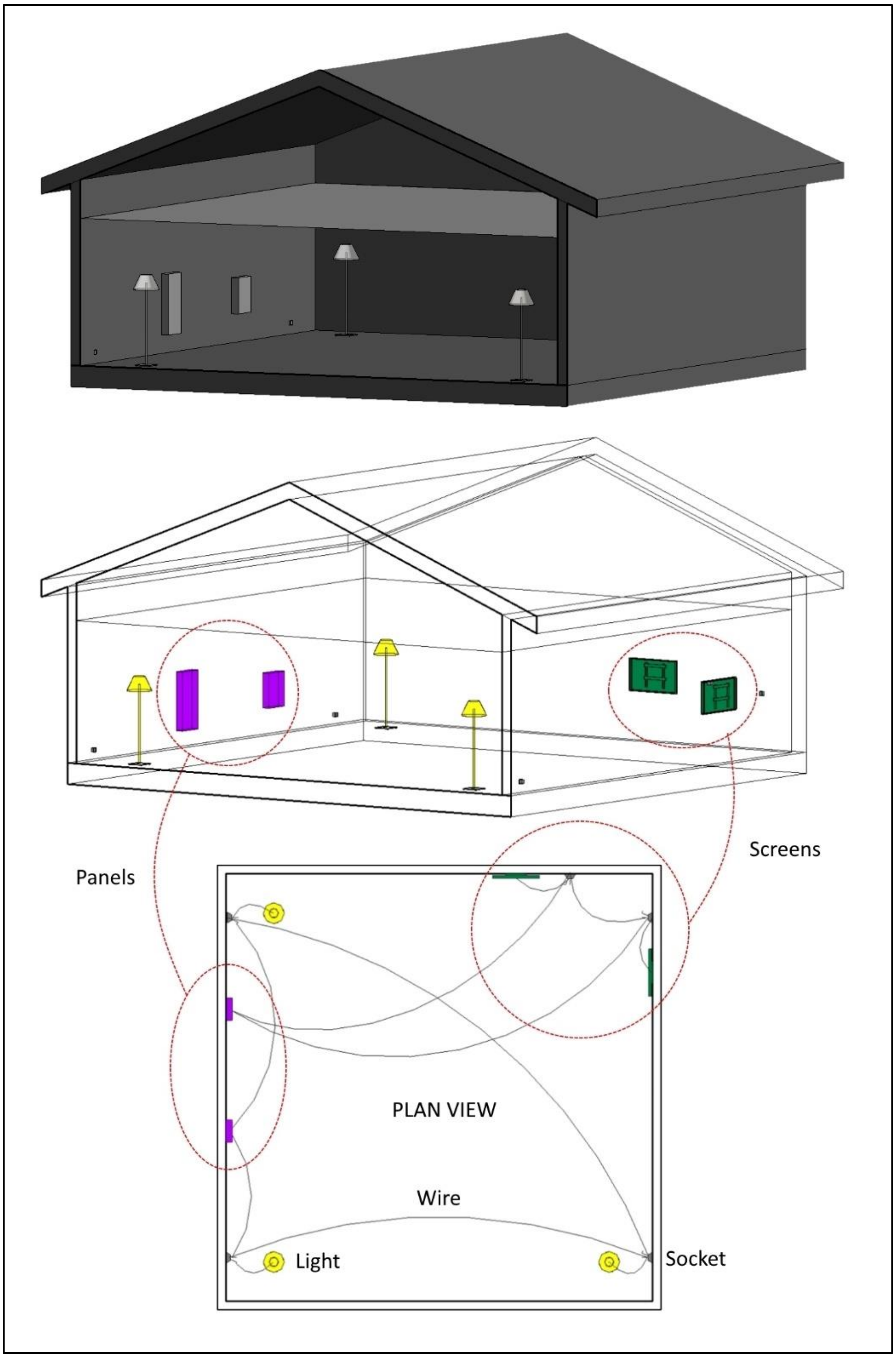

Figure 3.3.3-1 - BIM model of the low-complexity, synthetic Building X with lights, screens, electric panels, electric sockets, and the cables that connect these elements.

Within the Revit modelling environment, the components were connected to form electrical circuits. Two ring circuits were created: one for the lighting and one for the screens. Wires branch off from the sockets on the circuits to feed the lights and screens, and each circuit is connected to electrical panels, which are treated as connection points into the building.

The Revit families for the electrical components were modified by adding IfcExportAs and IfcExportType parameters that specify the correspondence between the families and IFC elements. This ensured that the exported IFC model contained entities that are valid within the IFC2x3 schema and represent as accurately as possible the electrical components in the Revit model. Floor lights were exported as an IfcLightFixture (POINTSOURCE type), TV display screens as IfcElectricAppliance (TV type), wall sockets as IfcOutlet (POWEROUTLET type), and electric panels as IfcDistributionFlowElement (no type).

3.3.4 Real-time data streams

Despite the benefits of using a simple, manageable BIM model for the design and evaluation of an integration method, no real-world monitoring data can exist for a synthetic building (such as Building X) and emulators (or substitute real data) would need to be used instead. However, as described in section 3.1, the USB is a highly monitored building that contains approximately 4000 sensors, measuring variables such as air temperature, humidity, lighting level, occupancy, and electricity consumption. The USB's BMS produces messages that are represented in the Building Automation and Control network (BACnet) protocol. BACnet is an ISO standard and a national standard in more than 30 countries (including the UK) (ASHRAE, 2020), and is the "most widely used standard protocol for building automation" (KMC Controls, 2018). The Urban Observatory (UO) (NCL University, 2017) collects readings that indicate *changes* in measured values (within a predefined tolerance). Across the entire building, approximately 40 changes are recorded each second, including those to electricity usage of various types. Although the UO's interface is not standardised, the standardisation of BACnet ensures that the streaming capability offered by the UO is readily reproducible for most other buildings with a building management system (BMS). These changes are encoded in JavaScript Object Notation (JSON) and published via an Application Programming Interface (API) and a websocket.

The UO publishes its data openly and the support of relevant research projects is within its remit and interest. Support for the USB from NG Bailey continued beyond the construction phase and their site visits were used as opportunities for discussion aimed at better understanding the mechanism of consumer unit monitoring within the USB and to interpret message attribute data.

For example, it was necessary to clarify that the appearance of 'C1' in the identifier of a message indicates that it refers to resource consumption within Core 1 of the USB (see Figure 3.3.4-1).

The USB has six floors (including the ground floor but neither the basement nor roof) and each floor is zoned horizontally into three cores. A set of Revit projects and IFC models exist for the USB – one of the three 3D architectural IFC models is presented in Figure 3.3.4-1, along with a plan view of the zoning of the third floor into its three horizontal cores.

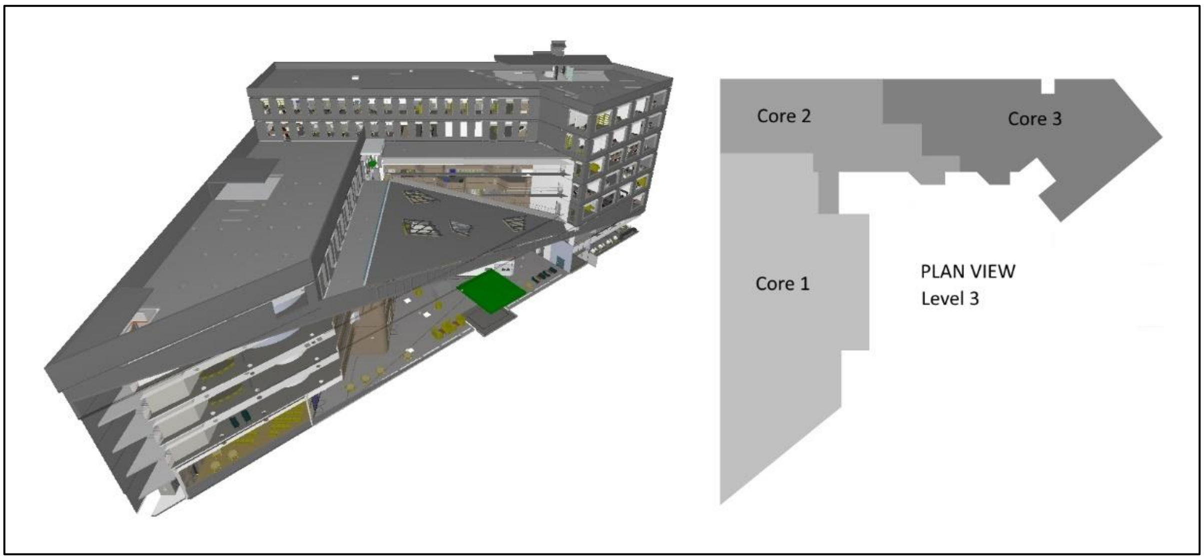

Figure 3.3.4-1 - A 3D architectural IFC model of the Urban Sciences Building, Newcastle upon Tyne, UK, and a plan view of the electrical supply zoning layout for the third floor.

Many of the electrical consumer units in the USB are serviced via one of three busbars (metallic strips or bars), which run vertically through the building, feeding subsidiary electrical networking and consumer units throughout the building. The data streams describing electricity consumption in the USB identify the type and quantity of usage, and the location of this usage within the building (the vertical floor and horizontal core). Figure 3.3.4-2 shows a single JSON message with attributes that describe the data recorded by a single sensor. The "id" value specifies the consumer type and core to which the message relates; the values of "unit" and "data" provide a real-time power figure for the consumer type; "buildingFloor" values identify the floor on which the consumer sits; and the "building" value identifies the USB.

```
{
  "RECEIVED": {
    "signal": 2,
    "data": {
      "brokerage": {
        "broker": {
          "id": "BMS-USB-1-JACE",
          "meta": {
            "protocol": "BACNET",
            "building": "Urban Sciences Building",
            "buildingFloor": "1"
          }
        },
        "id": "Drivers.C3_Electric_Meters.C3_Lighting_DB.points.C3_LightingDB_Power",
        "meta": {}
      },
      "entity": {
        "name": "Urban Sciences Building: Floor 1",
        "meta": {
          "building": "Urban Sciences Building",
          "buildingFloor": "1"
        }
      },
      "feed": {
        "metric": "C3 LightingDB Power",
        "meta": {}
      },
      "timeseries": {
        "unit": "kilowatts",
        "value": {
          "time": "2018-04-10T10:22:59.254Z",
          "timeAccuracy": 4.318,
          "data": 1.450948715209961,
          "type": "Real"
        }
      }
    },
    "recipients": 0
  }
}
```


The value of the "id" key is used to identify uniquely the consumer type and core; for this instance, the value describes the lighting power consumption in Core 3.

The values of the "building" and "buildingFloor" keys are used to identify the building and floor to which the message relates.

The values of the "unit" and "data" keys for the consumer types are used to populate the relationship power properties.

Figure 3.3.4-2 - A single JSON message recording the lighting power consumption in core 3 of the first floor of the USB as 1.45 kilowatts.

The simple key-value pair structure of JSON and the human-readable attribute names allow the messages to be filtered for a manageable subset that can be handled in this study despite the variety and velocity of data from the websocket.

3.4 Method

3.4.1 Method overview

Chapter 2 highlighted that the concept of a building overlaps the thematic remit of both the geospatial and BIM domains, and both of the CityGML and IFC standards (reviewed in section 2.3.4). Despite some disparity in the decomposition a building by each standard, the use of 'building' as an object name and the interpretation of the meaning of this name is identical in these two standards. The buildings represented in the IFC, CityGML and JSON datasets described in section 3.3 all reference buildings commonly. A method of integrating these datasets that exploits the commonality of referencing building entities is now explored.

Beyond this use of a common concept, the method is based on the principle that graph theory can support the understanding and analysis of urban spatial topologies and integrating models of urban data (Falkowski and Ebert, 2009; de Almeida, Morley and Dowman, 2013), and that graph databases can be used for efficient storage and querying of topologically connected data (Holzschuher and Peinl, 2013; Khan and Shahzad, 2017). The objective is to make best use of building identification (within urban datasets) and graph theory to derive an end-to-end electricity demand-supply network topology from the subnetworks represented by the disparate data sources. In support of the demand-side management use case that this chapter aims to address, also explored is the use of message brokerage technology for the dissemination of the dynamic state of an integrated demand-supply network. This latter part of the method is aimed at enabling a visual comprehension of resource flows as a means of managing electricity demand, which is also explored through the development of a web architecture.

An overview of the final software system that implements the method is represented in the flow chart of Figure 3.4.1-1. The subsequent sections of this chapter describe the incremental development of the method and software system around the case study datasets and in support of the target use case. As directed by the research approach of section 2.6, the software system was developed iteratively and subject to changing requirements as the datasets are explored and better understood throughout the study. Some of the key technologies that were chosen for the final system implementation are also shown in Figure 3.4.1-1.

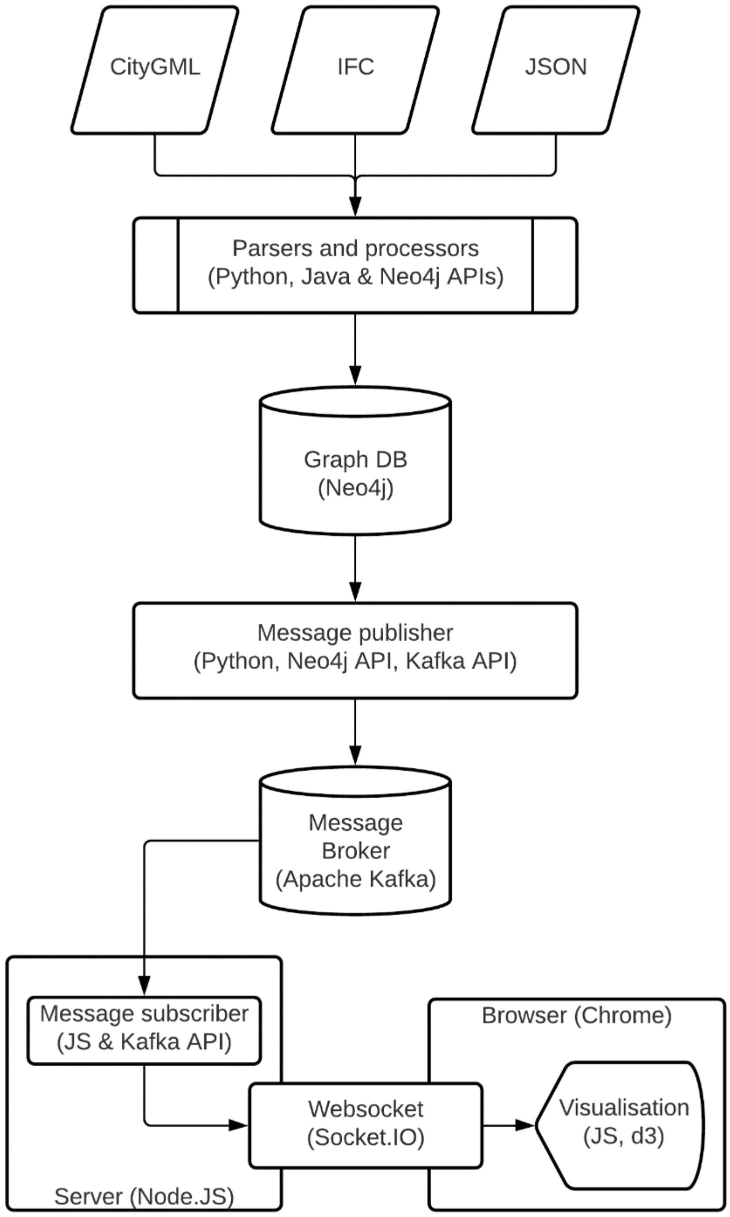

Figure 3.4.1-1 - A flow chart showing the method developed during this case study. The construction of this system is explained in detail throughout subsequent sections. The three data sources are parsed and process for salient elements, which are then pushed to a Neo4j graph database (DB) instance. The datasets are integrated within the DB, the state of which is published to a message broker for exploitation by a web server-socket-client visualisation system. For the implementation shown later in this chapter, the CityGML and IFC data is integrated only once but the JSON data is integrated repeatedly.

3.4.2 Modification of the building identifiers

Although the datasets described in section 3.3 identify buildings uniquely, they do so with different identifiers (and with uniqueness of different scope) such that no direct correspondence can be made. The building identifiers used in each dataset needed to be modified in order that the datasets could exploit the common use of a building concept for integration.

The footprint in the geospatial dataset (section 3.3.2) for Building X has a Great Britain (GB) Ordnance Survey topographic identifier (TOID), which identifies it uniquely within GB. The other building node represents Newcastle University's Urban Sciences Building (USB). Having been completed in September 2017 (Newcastle University, 2018b), the USB was too new for representation in Ordnance Survey data at the time this case study was carried out (2017); for this reason, the node was manually assigned an artificial identifier (in 2018, the USB was attributed TOID 5000005215818799).

Revit allows the assignment of a name to the project in the 'Building Name' parameter, the value of which is automatically attributed to the IfcBuilding element in the exported IFC version of the model. This name thus becomes the identifier for the building. Given the intended use of building identification in the integration process and the known OS TOID for the footprint to which BX is assigned, the name of the Revit project was also assigned this TOID value. However, there was no enforcement of format or uniqueness for the building name; the matching of identifiers between IFC model and building footprint was a design choice and had to be carried out manually.

3.4.3 Extraction and integration of the network topology

A Python script was written to extract internal building network elements and their topology from the IFC model using the IfcOpenShell-python module (IfcOpenShell contributors, 2020). IfcFlowTerminal, IfcFlowController and IfcDistributionFlowElement elements are extracted from the IFC file with their attributes identifying them as light and screens (the flow terminals), switches (the flow controllers) and distribution panels (the distribution flow elements). IfcDistributionPort, IfcRelConnectsPortToElement and IfcRelConnectsPorts relationships are used to connect these elements to each other (see Figure 3.4.3-1). The IfcBuilding element's 'Building Name' attribute identifies the building described by the IFC model.

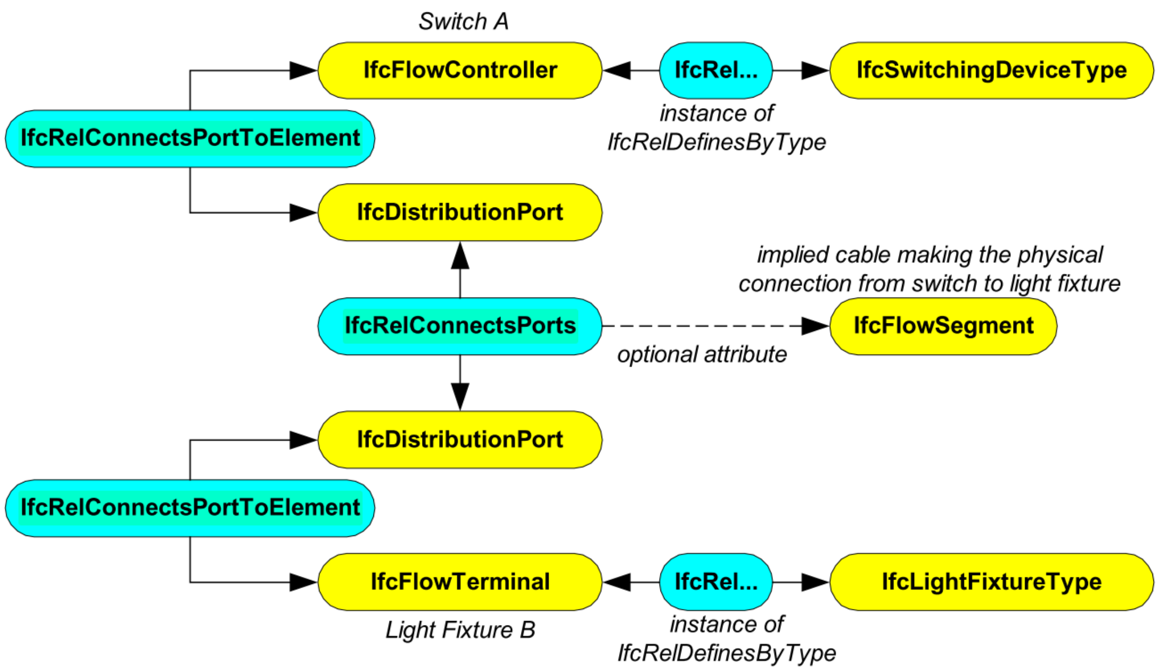

Figure 3.4.3-1 - IFC elements used for representing utility network topology. This figure is taken from the IFC Model Implementation Guide (BuildingSMART and Leibich, 2009) written by Thomas Liebich. The diagram shows how various relationship elements (such as IfcRelConnectsPorts) can be used to connect entity elements (such IfcFlowTerminal).

For the CityGML data, a custom Document Object Model (DOM) parser was developed to traverse the XML tree, extract Node elements from a FeatureGraph and uses InteriorFeatureGraph XLink references to connect the nodes. Figure 3.3.2-2 show the structure of these CityGML elements.

Neo4j graph database (DB) was selected for graph network implementation due to its freely available community edition, simple query language (Cypher) and support for the Python language. The topology extracted by the parsers from the CityGML and IFC files is pushed to an instance of the Neo4J graph database using Python scripts and the Py2Neo library. The sub-networks derived from each data source are integrated into a single network by exploiting building entity references that are common across the CityGML and IFC files. Neo4j Cypher queries run ‘merge’ clauses on building nodes with matching values of attributes that uniquely identify the buildings to which the source data relate. Use of the merge clause (rather than the 'create' clause) avoids duplication by only creating a node if it does not already exist in the database. Given that the GML IDs for the BX in the CityGML file matches the name given to

IfcBuilding element in the IFC model for BX, the two subnetworks are joined via this common entity.

Figure 3.4.3-2 shows the integrated topology in Neo4j graph database. The figure shows the single substation, two buildings (Building X and the USB), the nodes at which the buildings and substation connect to the distribution network (access points), and the electrical panels, light fixtures and screens inside Building X. The integrated network is shown at a minimal level of abstraction, representing the complete spatial topology derived from the data sources. These nodes can be related directly to the elements in the distribution network and BIM model.

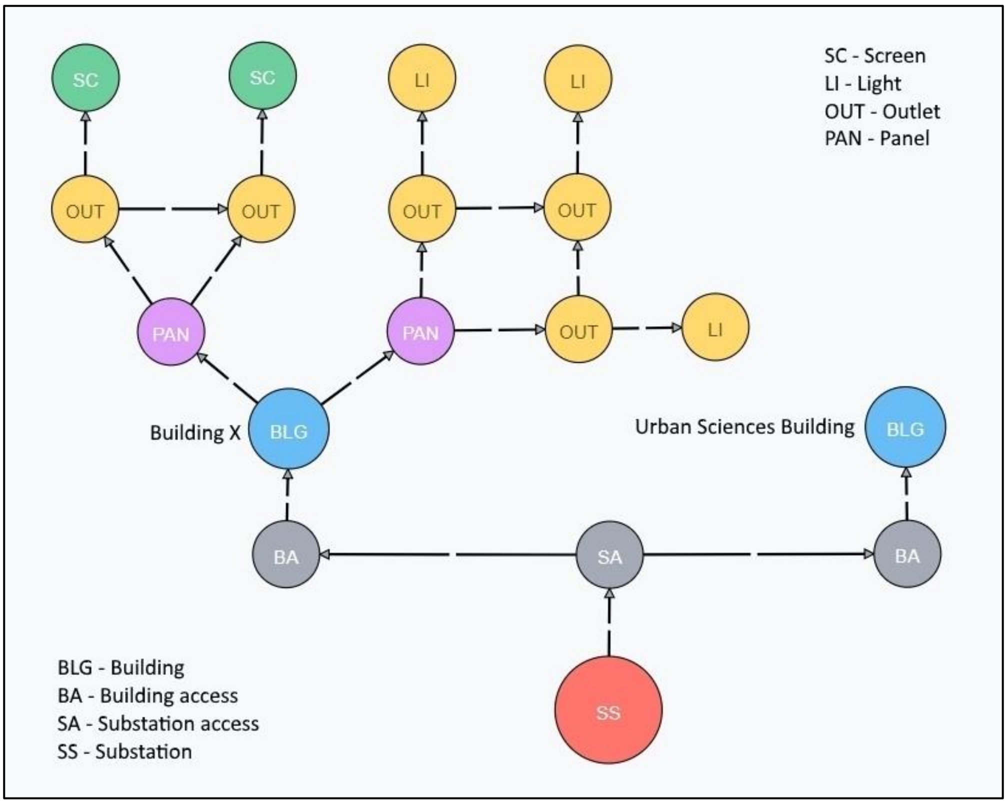

Figure 3.4.3-2 - A graph network representation of the integrated topology of the electricity distribution network and the internal electrical components of Building X.

Some infrastructure network modelling use cases – such as simulating the physical propagation of the effects of a network node failure – may demand a highly complex representation of elements, connections, and attributes. However, for an electricity demand-supply visualisation use, much of the complexity is redundant and a clearer understanding of resource flows may be presented through a simpler, abstracted topology. For example, a facilities manager might be interested in energy losses by identifying upstream supply (from a substation, for example) that is not accounted for by downstream consumption (within a building), but has no concern for the physical elements through which the electricity passes (see the use case of section 2.2.2). The Python script that integrates the CityGML and IFC subnetworks within Neo4j was modified such that it merges the nodes and relationships between the substation, buildings, and consumer nodes into single relationships; these new relationships connect the substation directly to the buildings and the buildings to their consumers. The hierarchy of this relatively simple, abstracted network topology represents changes in spatial scale without the complexity of connectivity between these layers.

3.4.4 Integration of the flow data with the structure of the network

The real-time data stream is now integrated with the CityGML and IFC topology in the Neo4j database. Whereas the UtilityNetwork ADE and IFC are domain-specific modelling schemas, JSON is an encoding language and is often implemented within schemas that are developed for specific technologies or use cases[33]. A JSON parser was developed specifically for the data streamed from the UO websocket. The Python script listens for messages from the socket and, on receipt of a new message, filters for those relating to three types of electricity usage: lighting, mechanical equipment and power sockets. The additional spatial information embedded in the message, which locates each message to a floor and horizontal core within the USB (see Figure 3.3.4-2), is also added to the graph network: the values of the floor and core in the message triggers the creation of a node for that floor and another node for the core (where they don't already exist), a link between the two and a link from the core to the usage type. In this way, the internal spatial structure of the USB and the types of electrical consumption within each space is built up in graph representation. The resultant, abstracted graph representation of BX and the USB are presented in Figure 3.4.4-1.

[33] An example is CityJSON (CityJSON contributors, 2019), which has been developed as an alternative to CityGML

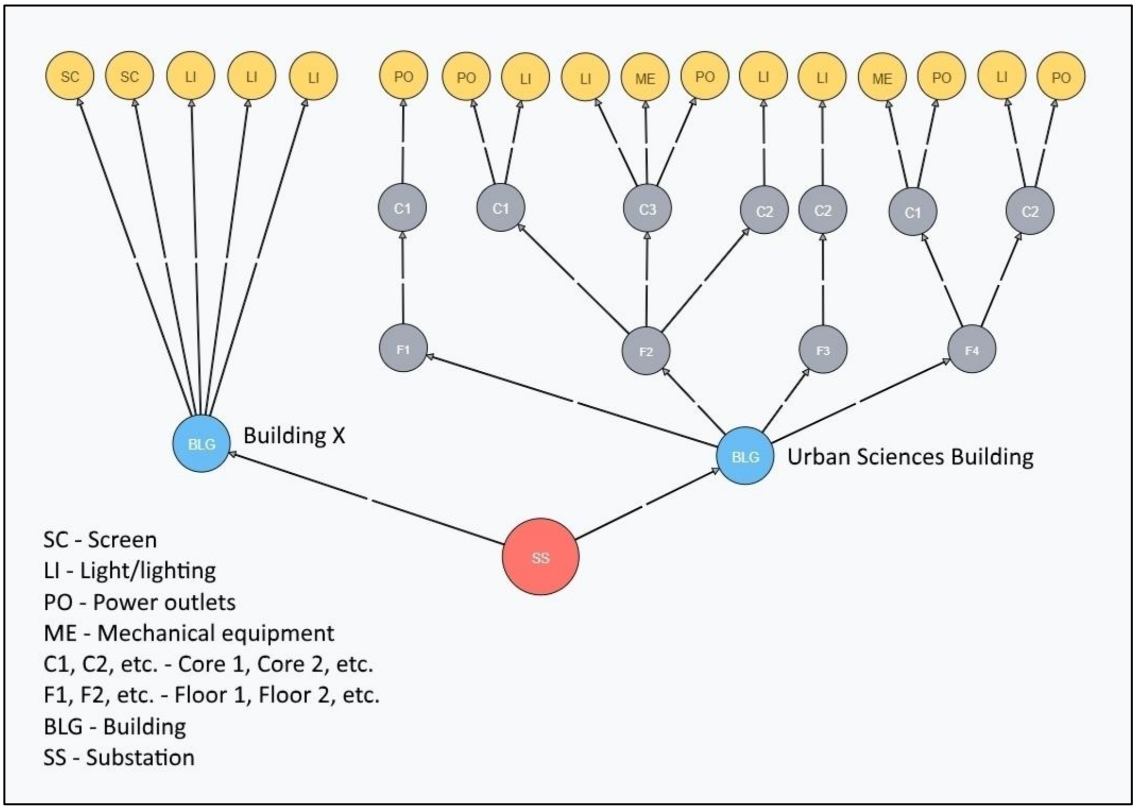

Figure 3.4.4-1 - An abstracted, integrated electricity network that represents types of electricity consumer units in each building. The network topology of Building X was abstracted to yield a simplified representation.

The spatial topology of electrical components inside the USB (shown in Figure 3.4.4-1) is derived from messages that were received over a period of approximately 5 seconds. Given exposure of the system to more messages, more consumer types within the USB will be discovered as they are referenced in the JSON messages. To demonstrate more clearly these dynamics, this subset of the network that describes the internal structure of the USB is modelled in isolation. Figure 3.4.4-2 shows the evolution of the integrated network when the graph database is populated with data from messages over a time window of approximately 10–15 seconds, captured at three points in time. As more messages are received from the data stream, more consumer types are identified across the floors and cores of the USB. Any new nodes, edges and attribute value updates are extracted from the JSON and pushed to the database via execution of merge queries. In Figure 3.4.4-1 (a), after exposure to the data stream for two or three seconds, messages have been received for two cores on the ground floor but only one core and one consumer type on each of the second, third and fourth floors. Around five seconds later, at (b), values have been received for other cores on the higher floors and for different consumer types. Given exposure to a further five seconds of sensor messages, at (c), the network is becoming even more populated by consumer types across the vertical floors and horizontal cores.

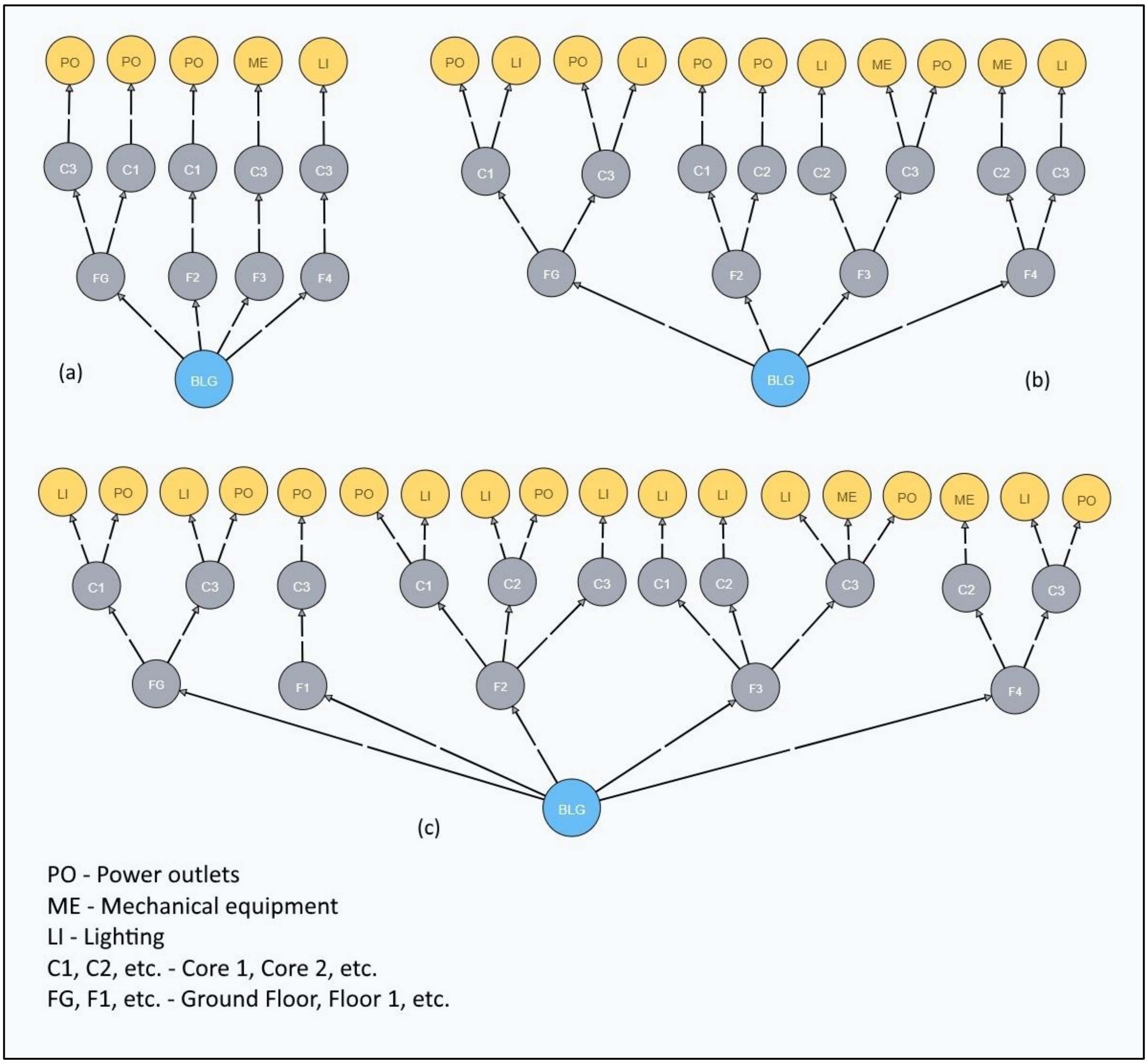

Figure 3.4.4-2 - An evolution of the graph network when the database is used to model only the topology derived for the USB, showing its growth from state (a) to (b) and then (c) as more messages are received from data stream. Any new nodes, edges and attribute value updates are extracted from the JSON messages received from the websocket and pushed to the graph database via execution of Cypher merge queries.

Beyond a population of the graph network with nodes representing floors, cores and electricity consumer types, the time series data and unit values from the USB data stream is used to assign real-time power values to an attribute on the graph relationships (links) that connect the core nodes to the consumer type nodes. This dynamic updating of a power attribute on the link is implemented through the same processing script that updates the structure of the graph network.

Throughout the hierarchy, the power values on each level are summed to provide a value for its parent level in the demand-supply tree. In the absence of any streaming data for BX, the individual lights and screens in the building are assigned constant, nominal values that are comparable in magnitude to those for the consumer types (which each represent the power used by multiple entities) in the USB.

On a technical implementation level, ‘set’ clauses are created from the 'unit' and 'data' attributes under the 'timeseries' and 'value' keys in the JSON messages (see Figure 3.3.4-2). These clauses form part of Cypher queries (written in Python) that update in real-time the values on an attribute of the links that connect to the target consumer nodes. The links further up the demand-supply tree are also updated with set clauses.

This real-time population of the graph database with new nodes and attribution of power consumption values to links constitutes a dynamic simulation of resource flows through the infrastructure network within the graph database.

3.4.5 Communication of the evolution state of the integrated network

In order to experiment with the practical usefulness of the dynamic graph representation of the network with respect to the target use case, structure and real-time flow-state of the network need to be made available for exploitation by potential end users. The evolving structure and state of the integrated network can be represented in time series by repeatedly producing JSON that represents the current, real-time structure (nodes and links) and state (values of the links' power attribute) of the electricity network. Each JSON representation of the network can then be communicated as a message.

Message brokers are well established software technology that handle streams of messages. There are two types of message broker models: point-to-point and publish/subscribe. Point-to-point models are based on messages residing in a queue for consumption once by a single user; in publish/subscribe models, messages are effectively broadcast through a topic from which multiple users may consume (IBM, 2021). The Apache Kafka streaming platform (Apache Foundation, 2020) offers a publish/subscribe brokerage capability (serving also as a message storage system), making it suitable for exposing the state of the integrated network to multiple potential exploitation systems. The distributed, scalable, elastic and fault-tolerant functionality of Kafka (Apache Foundation, 2020) also ensures that the system developed for this study is not limited in these respects by the brokerage software. Figure 3.4.5-1 shows a producer writing a message to a topic on a Kafka instance and two consumers reading from this topic. Consumers read independently from the topic with their own offset, which is record of the position of the consumer within the topic. The offset position can be set by the consumer, such that the stream of message can be replayed from anywhere within the topic. Although the implementation in this case study uses only one consumer, this multi-consumer characteristic is important for the scalability of the method.

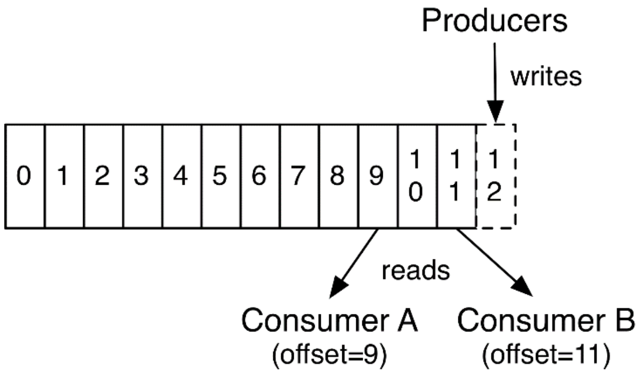

Figure 3.4.5-1 - Apache Kafka topic construction. A producer writes to a topic and then two consumers read from the same topic independently from the position of their own offsets. The offsets are parameters that determine from where in the message stream the consumers read. Image source: https://kafka.apache.org/intro.

In order to capture the real-time state of the network, a daemon (background application) recursively executes Cypher queries (at a defined time-step) that capture the state of the entire Neo4j database instance and store this state as JSON. For this case study, the time interval for the recursion was set at a single second. Each JSON message contains an array of link objects; each of these is attributed a start node, end node and power value. These 'snapshots' are then published as messages to a topic on the Kafka instance. Any other systems may then connect to the broker, receive the messages by subscribing to the topic and exploit them for analysis or

. Kafka also allows for outputs from other processing and modelling workflows to be published as messages to the same or another topic on the same broker. The Kafka topic to which the network state messages are published serves as a broker and store of time series messages describing the real-time evolution of the integrated electricity network.

3.4.6 Visualisation of the evolving network

A web server-client demonstrator system was developed as a means of evaluating the effectiveness of this brokerage method against the use case of electricity demand-supply visualisation. The components of this system and technologies used are depicted at the bottom of Figure 3.4.1-1. A Node.JS web server is deployed with a script that subscribes to the relevant topic on the broker and sends the received JSON messages through another web socket (developed using the Socket.io library) to a connecting web browser (Chrome was used in this study). In conjunction with JavaScript visualisation code and the HTML provided by the server, the browser uses the messages to display a dynamic Sankey diagram. The visualisation script is based on the d3 and d3-sankey JavaScript libraries (Bostock, 2018, 2019). The source code developed for these visualisation (and other implementations of the method developed for this chapter) are available at the Git repositories described in the Impact Statement of this thesis.

Figure 3.4.6-1 shows an example visualisation that represents the electricity consumption from the substation to the two individual buildings (the USB and Building X) and through to the individual consumers within the buildings. The vertical bars represent the network nodes and the connecting grey curved bands represent the relationships between these nodes; this structure relates directly to topology depicted in Figure 3.4.4-1 (the colours correspond). The heights of the bars and widths of the bands correspond to the relative magnitude of power consumption between nodes. The individual consumer types in the USB were measured to consume electricity in the range 0.1–10 kilowatts; for Building X, the consumer elements were assigned a constant 2-kilowatt output.

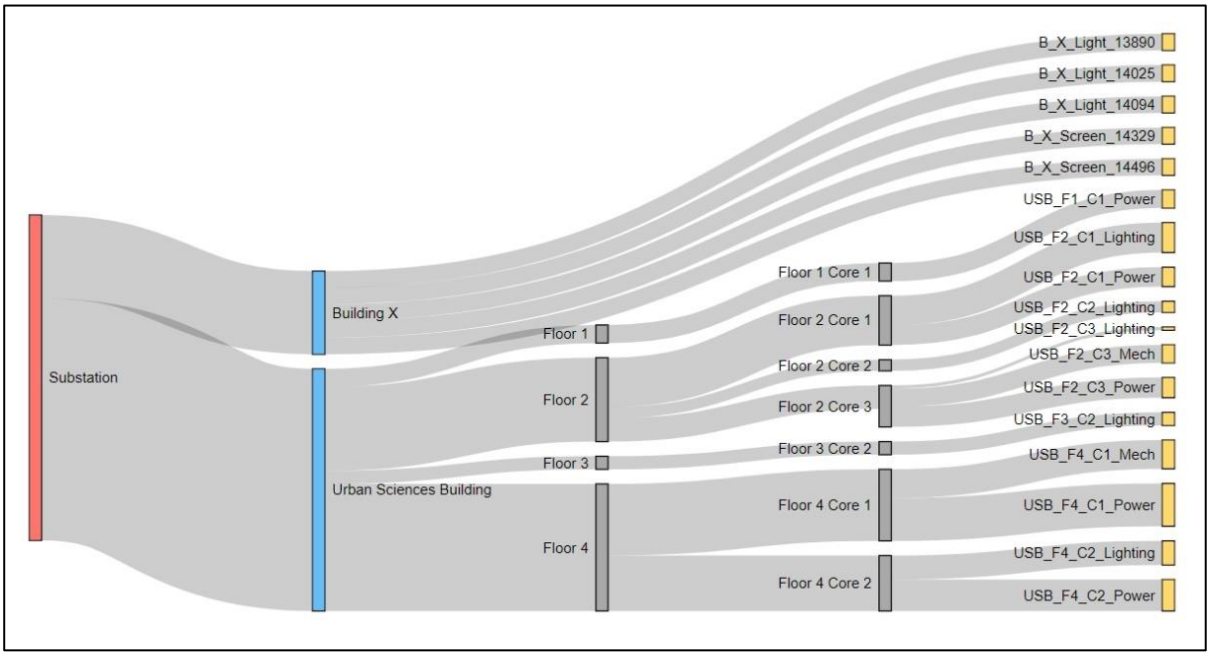

Figure 3.4.6-1 - Screenshot of a dynamic Sankey diagram, showing electrical power consumption through the network depicted in Figure 7. The thickness of the lines is proportional to the power consumption.

In the same way that Figure 3.4.6-1 shows the flow of electricity through the network of Figure 3.4.4-1, the Sankey diagrams of Figure 3.4.6-2 corresponds directly to the evolving network of Figure 3.4.4-2 for the USB in isolation. With reference to Figure 3.4.6-2: by the time that the network has reached state (c), it can already be seen that Core 1 has a relatively high consumption across all of the floors (where data for Core 1 have been received); it is also clear that the lighting (labelled ‘Lighting’) across the entire building is consuming more power than that the mechanical equipment (labelled ‘Mech’) and consumers that are connected to power sockets (labelled ‘Power’). The shrinking of vertical bar height for the USB from (a) to (c) is only a result of the visualisation needing to accommodate an increasing number of consumer types (and the spaces between them) on the right-hand side.

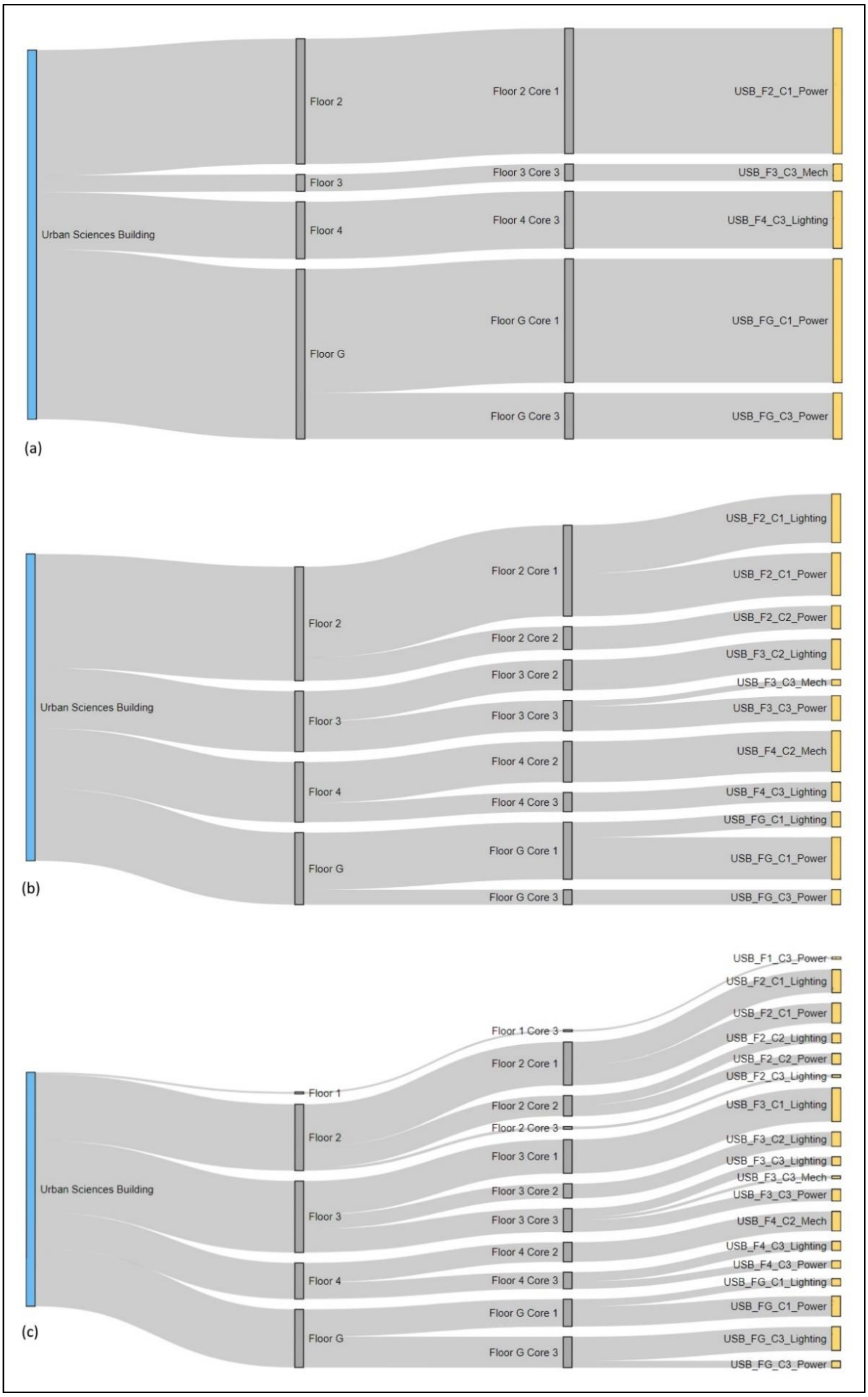

Figure 3.4.6-2 - Three snapshots, with a time-lapse of approximately 5 seconds, of an evolving visualisation of the flow of electricity through the Urban Science Building from state (a) to (b) and then (c); in real-time, the visualisation is updated each second.

3.5 Discussion of integration method

The method developed around the case study of this chapter and its datasets demonstrates the applicability of graph representations for continuous, real-time integration of the elements of datasets that span a range of spatial scales and are underpinned by heterogeneous GIS and BIM standards. Key to the method is the exploitation of reference to common concepts, consistent identification of real-world instances of these concepts and abstraction of detail such that only the information necessary for the target use case is represented. The concept of a building lies at the interface of the two domains. Disjoint digital representations can be joined by merging building features when they were identified as referring to the same object. By filtering out geometric, geographic and semantic detail, flow through the integrated topologies were directly comparable within the visualisation. However, an important shortfall of merging by object identifiers concerns the scope of uniqueness and commonality of identification. Firstly, only the OS footprint TOIDs are globally unique references. The name in the Revit project was set manually and the continued, confident use of 'Urban Sciences Building' as a building reference in the JSON data stream (see Figure 3.3.4-2) depends on no other building ever having the same name – this is not an assumption that can be made globally. Furthermore, even if every building reference was a globally unique identifier (GUID), there is no standardised way of ensuring that they match.

Graph databases are a useful technology for integrating and analysing urban data. Despite the concept of a building being common to the datasets of this study, much of the other data structures, concepts and semantics in the topological hierarchies are dissimilar. For example, while the CityGML UtilityNetwork ADE uses InteriorFeatureLink to represent connections between features, and IFC uses IfcRelConnects elements; and the identification of horizontal building cores in the JSON data could be represented as 'spaces' in IFC but there no known equivalent in CityGML (it would be inaccurate to represent each core as a 'room'). This potential problem is mitigated, to some extent, by the flexibility of graph database schemas: there is no need to predefine the database schema since new nodes, relationships, and properties can be added on-the-fly. As mentioned in section 2.4.6, this is consistent with the open-world assumption (OWA), which asserts that something may be true even if it is not known to be true. This freedom is powerful for the integration of the diverse and dynamic data sources encountered in urban data integration. Further to its suitability for modelling and processing urban topologies (Falkowski and Ebert, 2009; de Almeida, Morley and Dowman, 2013), a graph database enables intuitive, human-readable concept models of systems to be reflected in the database structure, which results in a schema that is often easier to interpret – this characteristic is shared with ontologies, which are recognised as being closer to a person's cognitive model that conceptual schemas (see section 2.4.2 for clarification of the difference). Any broader requirement for integration methods to enable deeper mathematical modelling (Song *et al.*, 2017) is facilitated by the speed and efficiency of using graph databases to execute queries on connected data (Holzschuher and Peinl, 2013; Khan and Shahzad, 2017), and the relative simplicity of constructing query statements.

The use of a message broker as a hub for storage and communication of messages describing network structure and flows proved effective for the target use case (section 2.2.2), and the technologies are highly scalable, but the scalability of the method that uses them in this study should be considered. The subscription to a topic is equivalent to subscribing to the database query that feeds that topic, with the topic serving as a view of the graph database that is tailored to the requirements of the use case. The Kafka and Neo4j technologies used in this study is highly scalable (Apache Foundation, 2020; Neo4j, 2021); Kafka, for example, can be deployed as a cluster spanning several servers, with the ability to communicate data through multiple topics for multiple use cases, and with multiple consumers subscribed to a topic. By using a websocket to publish topic messages, updates are *pushed* to the clients – there is no need for a request. The complexity and size of urban data models is increasing, as is the velocity and variability of real-time data from urban instrumentation. Although the technologies used in this study are efficiently scalable, the repeated and high frequency capturing of snapshots of the entire state of a large, complex network may result in a high-volume data stream that is largely redundant. Instead, the messages could capture *changes* to the state of a network, from which a client system could then reconstruct complete topologies and flows from multiple messages stored on the broker.

A more refined and scalable version of the system developed in this study has the potential to provide a basis for various other urban data integration and spatiotemporal analysis requirements. For example, a regional energy provider could deploy the components of the system that perform the integration and dissemination (based around the graph database and message broker), publishing messages that describe the state of the network that it manages. A facility manager could then subscribe to the relevant topic, exploiting the published messages on a dashboard that supports fault diagnosis through highlighting anomalies in a flow visualisation; at the same time, an urban planner could also subscribe to the topic and conduct analyses on historic time-series data in support of assessing the impact of proposed modifications to the supporting infrastructure. The geometries and geolocations encoded in the data sources could be used for several other purposes: to render multi-scale 3D visualisation in augmented or virtual reality environments, providing a more intuitive and immersive visualisation platform to users (Wang, 2009; Chi, Kang and Wang, 2013); for the detection of clashes between physical assets that would otherwise be represented in disjoint GIS and BIM models; or the ability to drill down from an urban-scale topographic map to view the real-time electricity flow within an individual dwelling or factory. The utilisation of timestamps on messages in the broker topic (or on updates to the graph database) would allow digital playbacks of network flow evolution from historic time series data. The system could facilitate the diagnosis of anomalies in usage patterns with alerts issued for values falling outside a predefined tolerance—for example, identifying that an increase in energy costs is due to the machinery in a factory activating erroneously at night.

Future work could attempt to verify that a version of the method developed in this study could be extended to networks carrying other types of resource, such as gas and water. The modelling of multi-resource interdependencies could show visually the effects of these couplings on flow dynamics. Such a system would also enable analyses for purposes such as root cause analysis. For example, if a set of resource provision services have failed, querying a multi-resource graph representation of the networks for upstream nodes that are parents (functionally and hierarchically) of all failed service end-points (a common dependency), a potential root cause can be identified (Neo4j, 2021). Erdener *et al.* (2014) identify gas-fired power plants in electricity systems and electricity-driven compressors in gas systems as the most significant dependencies in coupled gas-electricity systems, and active demand-side response (DSR) strategies stand to benefit from the modelling of coupled systems; Qadrdan et al. (Qadrdan *et al.*, 2017) showed that a significant reduction in gas consumption can be achieved by electricity peak shaving through DSR.

As is described in section 3.4.2 and the beginning of this discussion, the building identifiers used in each dataset needed to be modified to enable integration in this case study. One of the research gaps identified in section 2.5 is the use of real-world positioning for integration of network topology given the close relationship between this and the spatial topology (defined in footnote 22) of geospatially constrained utility networks. The following chapter tackles this research topic in the context of potable water supply networks and involve more interaction with data producers and owners, and application domain experts as specified by the research approach of section 2.6.

3.6 Summary of Chapter 3

This chapter described the development of a data integration method and implementation system that targets the use case of electricity demand-side management. The Urban Sciences Building (USB) and Helix site were used as a case study and a real-time dynamic Sankey visualisation demonstrates applicability of the method and system to the use case. The novelty of this chapter is, firstly, the use of representations of (or references to) buildings that feature commonly across multiple datasets for merging of digitally disjoint network topologies within a graph database, and, secondly, the integration of this graph database with a message broker and websockets for real-time monitoring of electricity flows across the multiple scales spanned by the datasets. The contribution is the demonstration that existing integration methods can be furthered by harnessing small data commonalities in otherwise disparate datasets with the support of flexible software technology.

The research showed the effectiveness of using graph representations and a graph database for the integration of disparate and disjoint urban datasets, and for the representation in real-time of electricity consumption across multiple spatial scales. The developed system also showed how publish/subscribe message brokers are an effective means of communicating dynamic network structures and flow states to multiple consumer systems. It is suggested that the use of common and universally unique identifiers is an important part of a solution to some integration challenges. The ability to assert correspondence between data streams output by sensors and the digital representations of their subject real-world assets or environments remains a significant challenge to the modelling of urban environments. The potential of using spatial data to infer connections between digitally disjoint infrastructure assets is identified as a research opportunity to be addressed in Chapter 4.

Chapter 4 Integration of potable water supply network topology across the building envelope

4.1 Introduction

The second and third use cases detailed in Chapter 2 (sections 2.2.3 and 2.2.4) concern water network partitioning and network configuration planning. This chapter addresses both of these use cases (the first in section 4.2; the second in 4.3) through a technical study of the water networking in and around three of the buildings of the Helix site, which was introduced at the beginning of Chapter 3 (refer to section 3.2 for details). This second case study makes more extensive use of the BIM MEP modelling available for the Helix site and, whereas Chapter 3 used a heuristically derived electricity distribution network data, this study makes use of a real-world dataset provided by the local water distribution network (WDN) operator, Northumbrian Water Ltd (NWL).

Due to the importance of gaining insights from NWL and access to the infrastructure datasets that are critical to the research of this chapter, and thus the need to provide sufficient incentive for engagement in discussions and sharing of data, it was necessary to consider the potential positive impact on the research, development and operations of NWL. In additional to the other stakeholder listed in section 1.4, remote meetings were held with the NWL Research & Development Manager and Continuous Improvement Lead to establish the factors that could be tackled by this research. The Continuous Improvement Lead provided a description of existing data-centric efforts underway within the company and discussed the types of analyses within this research that could enhance NWL's capabilities. These technical discussions influenced the research gaps of section 2.5, research questions and objectives of section 1.4, and use cases of section 2.2. While there was significant effort underway by NWL to integrate their geospatial and hydraulic datasets, the use of spatial heuristics and graph network theory for improvement of their infrastructure planning and optimisation were established as capability gaps worth pursuing.

Section 2.5 of the literature review identified the use of both object identification and real-world positioning for utility network integration as needing further research. While Chapter 3 investigated object identification, this chapter tackles the use of location data and geometry to address some problems that cannot be solved using the method developed in Chapter 3. Specifically, this chapter considers the closeness of relationship between the network topology of geospatially constrained utility networks and their spatial topology, and how spatial relationships may be used to make inferences of connections between networks where the absence of complete representations prohibit deductions. The ability to infer inter-dataset connectivity is researched in the context of potable water supply networks.

Within England and Wales, potable water is provided to consumers by privately owned companies that operate water distribution networks (WDNs). WDNs comprise pipes and other assets that are spatially and topologically complex (Yazdani and Jeffrey, 2011; Torres *et al.*, 2017), predominantly underground and challenging to manage (Ofwat, 2015). The water companies collect, treat, distribute and monitor the supply of water to individual premises; once delivered through the boundary of these premises, the pipework and any losses through leakage become the responsibility of the property owners. The complete demand-supply network spans multiple spatial scales, from reservoirs that feed regions down to consumer units in buildings. NWL owns and operates the water distribution infrastructure in North East England, including the Helix site.

Available for use in this case study are a geospatial-scale WDN dataset from NWL, building footprint data from Ordnance Survey (OS)'s MasterMap® Topography and VectorMap Local layers, and BIM MEP models produced by NG Bailey and TGA Consulting Engineers for the buildings of the Helix site.

The research of Chapter 3 involved little interaction with data producers and owners, and application domain experts (an aspect of the research approach – see section 2.6). This research of this chapters involved consultation with NWL, engineering contractors for the Helix project, the Estates department of Newcastle University and was influenced by discussion within the Integrated Digital Built Environment (IDBE) working group.

4.2 Spatial inference method of integration for water networks

4.2.1 Introduction and method overview

Topological integration of the digital representations of multi-scale water networks is subject to the challenges presented by disparities in data standardisation (see Chapter 2). Addressing research gaps identified in Chapter 2, this chapter examines the use of georeferences in the Helix buildings' BIM Mechanical, Electrical and Plumbing (MEP) models for integration of the topology of the buildings' water networks their surrounding geospatial WDN, what semantic and geometric information needs to be retained and what can be abstracted to overcome disparities in the data standards that underpin the source data.

The technical objective of this chapter is to develop a semi-automated method of inferring connections between the Helix buildings and the WDN that minimises dependency on semantic information. The method is based on the probabilistic assignment of plausibility to asset pairings using heuristics: more specifically, the identification of assets in a WDN that are candidates for connecting with the mains water entry point of a building and the use of engineering principles to assign likelihood scores to these competing candidates.

4.2.2 Geospatial and BIM data sources

Location and attribute data for the NWL distribution assets (pipes, valves and other components) are provided in an Esri geodatabase. The data of interest within this database are those representing mains potable water pipes, which are represented relative to the OSGB36® coordinate system as linear features comprising one or more pipe segments. The source data is two-dimensional (2D); the features are given heights by draping them over an interpolation of a 0.25 m resolution LiDAR Composite Digital Terrain Model (DTM). The assets are then uniformly offset by negative 1.1 m, representing an approximate average of the minimum and maximum coverage limits under design guidance and self-lay requirements (Thames Water, 2015; Northumbrian Water, 2018; Water UK, 2018). This process results in a three-dimensional digital representation of the pipe assets.

Building footprints are understood here to mean the ground area contacted by a building (Diakite and Zlatanova, 2020) and, for this study, footprints are sourced from OS Topography and VectorMap® Local layers. The OS Topography layer provides lines and polygons that describe linear building features in 2D. These features are provided with heights (bringing them into three-dimensions) by draped over the same DEM that was used for the NWL data (above). Values of 'obstructing' and 'overhead' on the 'physicalPresence' attribute for line features need to be used for distinguishing polygon features representing orthographic projections of overhanging building sections from those representing footprints. This process is of importance because it constitutes a derivation (rather than direct sourcing) of footprints that align with the above definition of a footprint, which has implications on the automation of the method. In order to avoid unnecessary data processing, abstracted representations of buildings available in the OS VectorMap Local (VML) layer (instead of the Topography layer) are used as building footprints for buildings other than the subject buildings of B1, B2 and B3; the discrepancies between precise building footprints and the abstracted building outlines from VML for these peripheral structures are small enough to be negligible for this study. Footprints from the VML layer are also draped over a DEM to bring them into 3D. The abstracted footprints disregard whether the outline is overhanging or in contact with the ground, which is not important for the analysis of the peripheral buildings (those that are not one of subject buildings - B1, B2 and B3) described later in this chapter. Figure 4.2.2-1 shows footprints and water distribution pipes for the study area, plus the overhangs of the subject buildings.

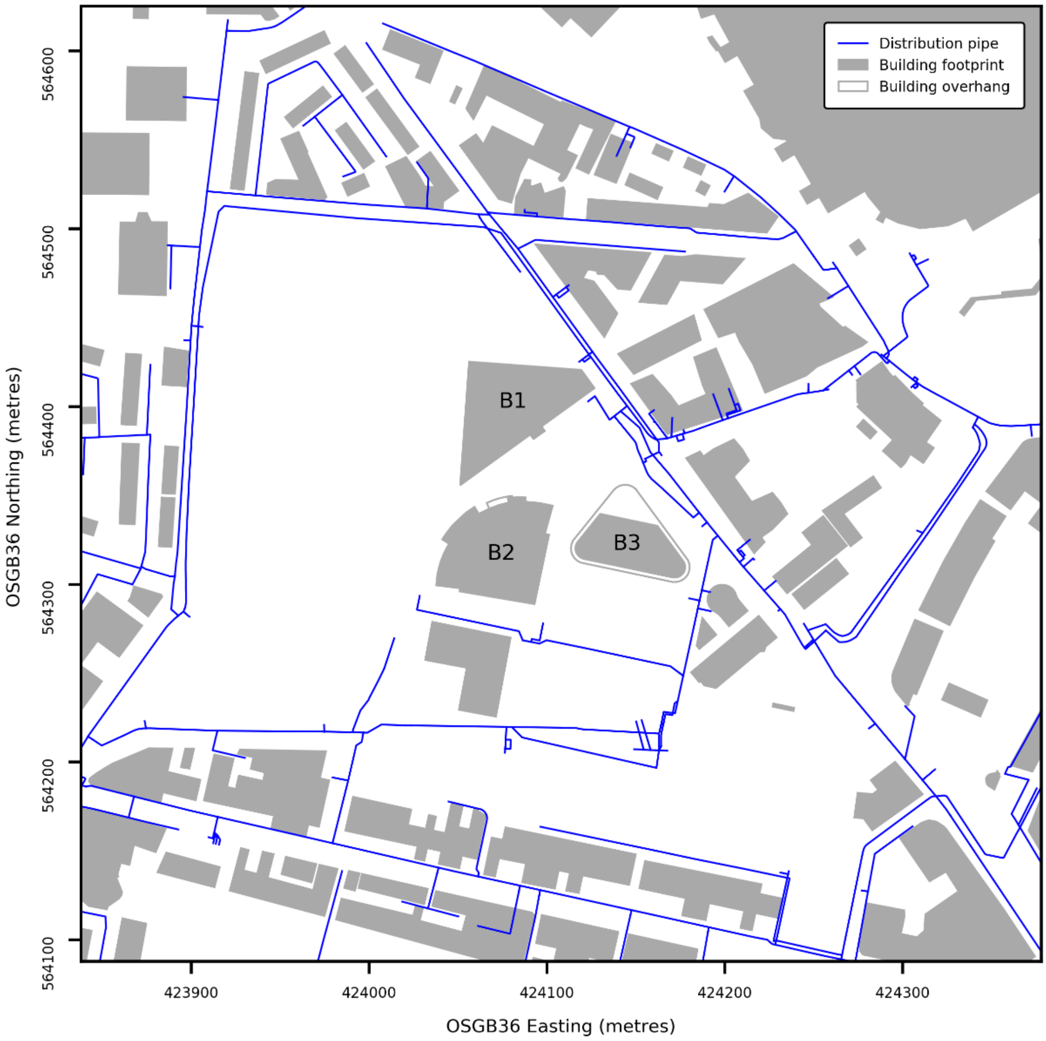

Figure 4.2.2-1 - Geospatial data used in this study (plan view). Building footprints are defined in this study as areas of the ground contacted by buildings. Buildings B1 (Urban Sciences Building), B2 (Frederick Douglass Centre) and B3 (Catalyst) are the subjects of this study. Note that the peripheral buildings (those that are not B1, B2 or B3) are represented at an abstracted level, without regard for whether the outlines are overhanding or in contact with the ground. Contains OS data © Crown copyright and database rights 2020 Ordnance Survey (100025252).

The footprints for the three subject buildings serve two purposes: firstly, they provide a means of checking the geospatial location of the Mechanical, Electrical and Plumbing (MEP) BIM models after transforming them from a local Cartesian coordinate reference system (CRS) into the geospatial OSBG36 CRS; secondly, they are representations of the building envelope that can be used for identifying BIM pipe assets that breach an external wall of a building and are thus likely water entry points into the building.

The BIM modelling of the site includes the MEP elements that represent the distribution systems inside the buildings. These MEP models are provided in the Industry Foundation

Classes (IFC) format (buildingSMART International, 2020a), including data describing network topology, geometry and geolocation. Although the models represent explicitly the water network topology between elements, in much of the source IFC data, correct flow direction is not preserved throughout entire flow chains, which has an implication for identifying flow terminals (this is discussed later). Pipe assets are represented in the IFC datasets as IfcFlowSegment elements using Swept Solid geometries. Connecting assets, such as IfcFlowController and IfcFlowFitting elements, are stored using the 'Boundary Representation' technique, sometimes within SolidModel, SurfaceModel and IfcFacetedBrep elements. All the geometries are with respect to a local engineering CRS and the Swept Solids make use of direction vectors (DirectionRatios) and depth values that indicate the axis (within the local CRS) and distance along which the pipe profiles should be swept. The IFC model also provides a single geolocation (in OSGB36®, through the project's IfcSite entity) and project vector that orientates the entire model with respect to true north. It is noteworthy that use of the OSGB36® CRS was not explicit in the BIM models and had to be assumed. The geolocation, project vector, asset geometries and a scaling factor for the local area (0.99960838) are used in a translation, rotation and scaling operation to convert the IFC water networking elements into the same geospatial CRS as the distribution assets and building footprints using custom Python scripts. Figure 4.2.2-2 shows a sample of some MEP elements from the IFC models (IfcFlowSegment elements) across all floors of the three buildings, some of which are part of the mains water network. The misalignment of the IFC assets with the building footprint for B3 in Figure 4.2.2-2 shows that the IFC model for B3 is not positioned correctly.

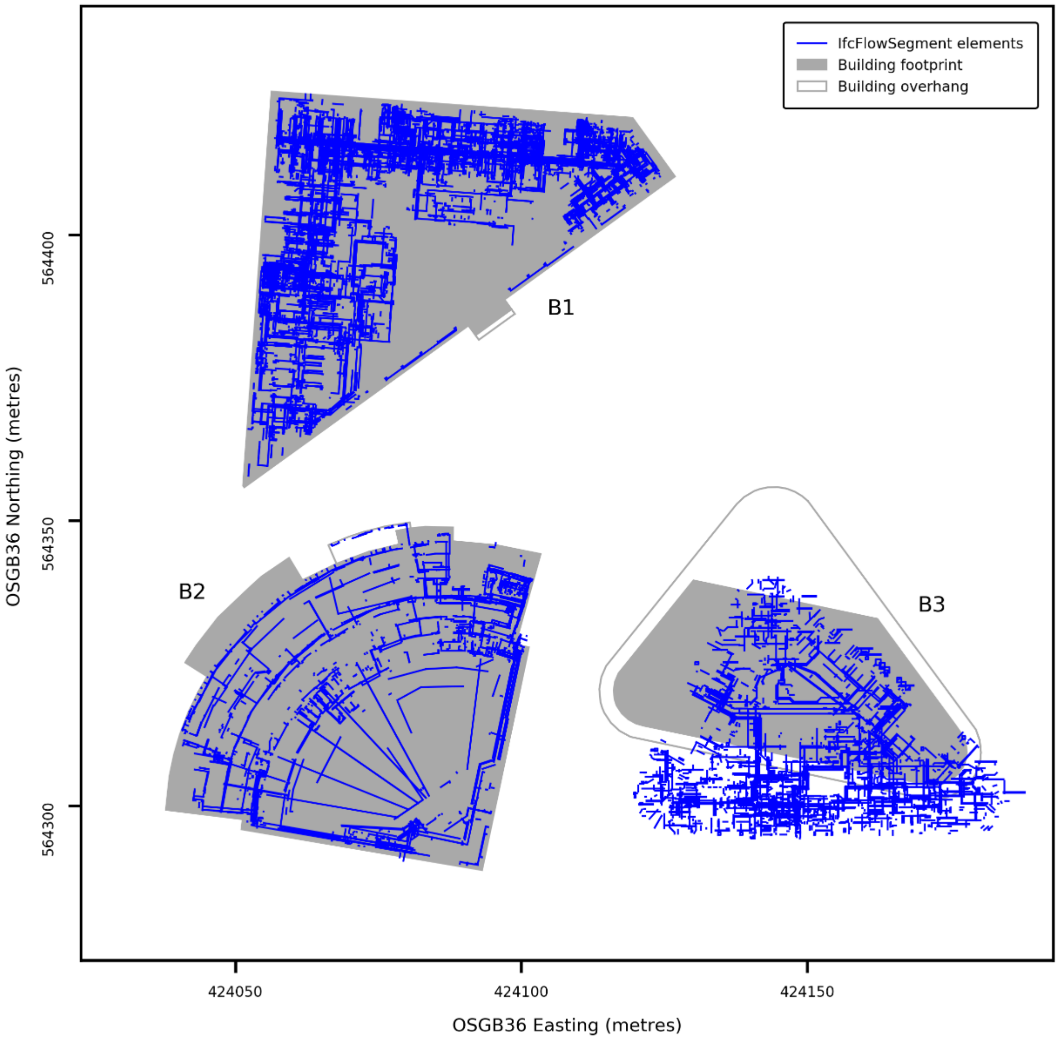

Figure 4.2.2-2 – Plan view of a sample of IfcFlowSegment elements (ducts, pipes and conduits, coloured in blue) across all floors of the three subject buildings; a subset of these elements are mains water pipes. Contains OS data © Crown copyright and database rights 2020 Ordnance Survey (100025252).

The BIM assets' clear alignment with (and almost entire containment within) the building footprints for B1 and B2 (see Figure 4.2.2-2) indicate that the BIM assets are positioned with at least a similar accuracy to the building footprints (as previously mentioned, the positions of the BIM assets for B3 were corrected manually). Direct surveying of the interior BIM main water assets or a measurement of their offsets from directly measurable features would be needed for a more confident assessment. This misalignment of the BIM assets for B3 is rectified manually by rotating the project vector for its IFC model by 11 degrees clockwise. The elevation value given in for B3's BIM model was corrected from 8.2 to 60.1 metres using the same draping method (and DTM) as was used for the NWL distribution assets. This need for a correction to the geolocation and orientation of B3, along with the need to assume the use of the OSGB36® CRS for all BIM models, is of particular importance due to the reliance of the positioning of all the BIM assets on these references.

The uncertainty in positioning of the data sources was measured by conducting a high-precision Global Navigation Satellite System (GNSS) survey of one of the building envelopes and several of the distribution assets, the data from which were then compared with the source data positions. The aim was to measure the positions of the features to the maximum accuracy that could be realistically achieved but with an uncertainty no higher than 50 mm (0.05 m).

One existing survey nail to the east of B1 served as an observation point for a Leica Geosystems MS60 Total Station (TS). A second nail was placed to the north of B1, enabling the orientation of the TS on the east nail using the single back-sight method. The east nail was positioned close enough for precise, unobstructed observation of building features but in a relatively dense urban environment; the north nail was positioned in a more open environment, farther from B1 (see Figure 4.2.2-3). It was ensured that the location of the north nail had line-of-sight to the east nail. The nails were observed over an approximate two-week period using a Leica Geosystems GS18 GNSS receiver. For each nail, five half-hour observations were carried out at different times of day between approximately 08:00 and 21:00 BST, ensuring coverage of a broad range of satellite constellations and enabling the calculation of a root-mean-square error (RMSE) for the observed position of each nail.

Figure 4.2.2-3 - Photographs of: [top] the GS18 GNSS receiver above the northern survey nail (N); [bottom-right] the GS18 above the eastern survey nail (E). [bottom-left] Google Satellite view showing the location of the survey nails with respect to the Urban Sciences Building (B1).

The RMSE of the position of the east nail was calculated as 3 millimetres for the Easting (424141.3536), 10 millimetres for the Northing (564399.3874) and 13 millimetres for elevation (63.9257); for the north nail, the RMSEs were 3 millimetres for the Easting (424055.5211), 2 millimetres for the Northing (564488.0892) and 4 millimetres (69.1895) for the elevation. Based on the RMSE data for the east nail, the maximum error in the XY (east-north) position of the TS is 11 millimetres (0.011 metres). The distance between the locations of the two nails was then calculated and, with the TS above the east nail and a prism placed above the north nail, the distance between the two nails was also measured directly. The difference between the calculated and directly measured distances was 2 millimetres, indicating that the 11 mm location error for each of the nails is a conservative estimate.

The two most easterly vertices of the building footprint of B1 were surveyed using the TS. Three additional vertices (above ground level) on the building straights were also measured at no more than 10m above ground level. Given the approximate verticality of the building straights and the proximity of the measured points to the ground, the XY of these measurements were used as approximations for points on the (ground contact) footprint. Direct measurement of the ground contact points was complicated by an absence of clean and consistent features. The measurements were used to verify that the OS MasterMap footprint for the USB was accurate to within 0.15 metres.

The accuracy measurements are consistent with what is expected of OS data. The OS surveyor responsible for mapping the Helix buildings was then shadowed during a re-surveying of parts of the site for changes; through observation of and discussion with the surveyor about the data collection techniques used for the Helix site, it is known that the surveying method and equipment used by OS for B1 were also used to produce footprints for B2 and B3; based on this, the uncertainty in position for the building footprint of B2 and B3 is assumed to be similar to that for B1. Through discussions with the surveyor and reference to OS documentation on positional accuracies (Ordnance Survey, 2020b), it is known that the RMSE for these footprints should be no more than 0.42m and 99% of points should fall within 0.9m of the recorded position.

The accuracy of the positions of the distribution network data were assessed by surveying the centre-points of seven manhole covers to the northeast, east and southeast of the USB. These covers were used as proxies for the position of the underground assets for which they provide access (valves and meter chambers). In the NE of England, pipes are usually be laid to a minimum cover of 900 mm and a maximum of 1350 (Northumbrian Water, 2018). The TS was used in conjunction with a surveying prism that was placed and held manually over each cover. It was estimated that an uncertainty of roughly 0.5 metres was introduced by the total of i) the approximation that these centroids represent the position of the underlying assets, ii) the estimation of the position of the centroid and iii) the manual holding of the prism in position during the measurements. To the nearest metre, the positions given returned by measurements varied from the positions in the dataset by up to 6 metres in Easting or Northing, an error that is more than an order of magnitude larger than those of the building footprint and BIM data; it is assumed that 6 metres is a reasonable approximation of the uncertainty in position of the distribution network assets as a whole.

4.2.3 Flow data collection and analysis

Flow data for the USB (B1) was sourced and analysed for this study as a means of verifying flow connectivity from the WDN to the building and to study the granularity of monitoring available to this case study. The FDC (B2) and Catalyst (B3) were not completed at the time of this study and thus their flow could not be studied. A significant amount of time was spent studying the connectivity of assets within the basement of the USB, asserting a correspondence between BMS data and the water meters (inside the basement) and understanding whether a combination of the connectivity and flow data could be used to identify and locate any leakages in the system.

On the urban scale, the flow of water into the DMA containing the USB is monitored by NWL at the boundary to the DMA and then at the boundary of the Helix site by the UO and Demeter Ltd (just before the water enters the connection pipe to the USB). The DMA flow data is not made available publicly in real-time but was requested and provided by NWL for the time window of 1st June to 21st August 2019. The boundary valve data is available in real-time via the UO's API and at https://www.checkyourwater.co.uk/ - access was provided to the researcher by Demeter for this study. Aside from the usage of a small subset of consumer units throughout the building, the USB mains water flow is monitored in several places in the basement: immediately after entry to the building, at the entries to the building's potable and CAT5 water tanks, and along two supply lines that provide water for sustainable drainage systems (SuDS) research facilities. Data for the mains water entry, potable and category 5 (CAT5) meters were collected from the UO API and via manual inspection of the analogue meters in the basement on a regular (approximately weekly) basis in August, September and October 2019.

Excluding the flow data available for the small subset of consumption units, the coverage flows no lower in the USB's water service hierarchy than the basement storage tanks prevented visualisations equivalent to those for generated for more granular electricity flows in Chapter 3; although the consumption data for the appliances of Building X (see section 3.3.3) were artificial and fixed, disaggregation of electricity usage into consumption types, floors and zones was derived from a real-time data stream.

The approximate matching of flow values for the mains water incomer to the USB and at the property boundary was a verification of the USB-WDN connectivity (see section 4.2.7). A significant mismatch would have indicated either that the USB connected at a different point in the WDN from what was expected or the presence of a leak in the pipes between the boundary and the building. However, the total flow values over ~2 months for of the potable, CAT5 and SuDS meter values were ~7% lower than for the mains water entry meter – the values should match given that there were no known leaks in the basement during the study period. This under-measurement of flow might be attributable to analogue metering sensitivities: it was noticed that the meter on one of the water tanks did not register consumption when water was trickling into the tank (very low flow rate).

4.2.4 Connection candidate selection process

The mains water pipes in the BIM MEP models are identified by selecting the subset of IFC elements that represent the mains water systems. Given that the BIM IFC model does not consistently preserve flow direction and thus cannot be used in isolation to identify the start of a water flow chain, the entry pipes to the building (in the BIM model) are identified by their approximately perpendicular intersection with the building envelope (a buffer of 0.2 m and angle tolerance of 5 degrees are applied).

The spatial topology of the distribution pipe assets in the geodatabase are used to identify assets that are candidates for connection to the building; those that do not touch another asset at one end are considered end-points and hence candidate water flow end points. It is recognised that this is not a full representation of reality – for example, some flow terminals feed fire hydrants and others represent false ends due to missing data. In order to test the effectiveness of the spatial-topological approach, a set of distribution candidates is also generated using a semantic filter that selects those assets whose 'NET_FUNC' (network function) attribute has the value

'PRIVATE' – this value indicates ownership of the asset by the consumer and provides a means of identifying semantically the connection points to the premises.

The subsequent aim is to develop a generic and transferable method for inferring connections between the distribution network (Figure 4.2.2-1) and BIM MEP networks (Figure 4.2.2-2) that does not rely on semantic information or conceptualisation. The inference method is heuristic and involves the probabilistic assignment of plausibility to asset pairings based on their spatial characteristics.

It has already been shown that a topological framework – instead of hydraulic simulations – can be used for preliminary WDN design purposes (Giudicianni *et al.*, 2018) and a similar approach here supports the potential for generalisation and transferability of the developed method to other utility resources. For reasons of practicality and cost, pipes that connect WDNs with serviced premises are preferably laid on shortest-path, direct transits with perpendicular joins and they should lie entirely within the serviced curtilage, remain accessible for repair and not be built upon (Amaral, Scala and Barthe, 2000; Atkinson, 2012; Thames Water, 2015; Affinity Water, 2018, 2019; Giudicianni *et al.*, 2018; Northumbrian Water, 2018; Yorkshire Water, 2018; Zhao, Liu and Mbachu, 2019). A deterministic and rule-based modelling approach can only account for all real-world scenarios if all of the factors influencing these scenarios are known and can be modelled; a probabilistic method, however, is able to attribute likelihood to possible connections without such strict dependencies. Based on the above guidelines and constraints, it is asserted that a water distribution network flow end-point is a more likely candidate if it is closer to the point of entry of water to the building (first criterion), if it points more towards it (second criterion) and if this pointing vector intersects less area of building footprints (third criterion). These three criteria form the basis of a probabilistic inference method that outputs likelihood scores for candidates. For existing network layouts, this likelihood may be considered equivalent to plausibility; for future builds, it can represent feasibility.

4.2.5 Algorithm development

An algorithm is developed around the above spatial inference approach. The distribution candidates are initially subset spatially to a bounding box within 200 m of the three buildings in the source BIM data, which is considered a reasonable approximate maximum Euclidean distance for connection transit (the selection outcome did not change for a wider spatial limit).

The distribution candidates then undergo a multi-criteria decision analysis (MCDA) in the form of a weighted sum model (WSM) for the three criteria. The WSM attributes and sums normalised values for each of the three criteria for each candidate, yielding a likelihood score. In the absence of evidence that would favour one criterion over another, the criteria are weighted equally. The most likely candidate is selected for pairing. The entire process is shown in the flow chart of Figure 4.2.5-1.

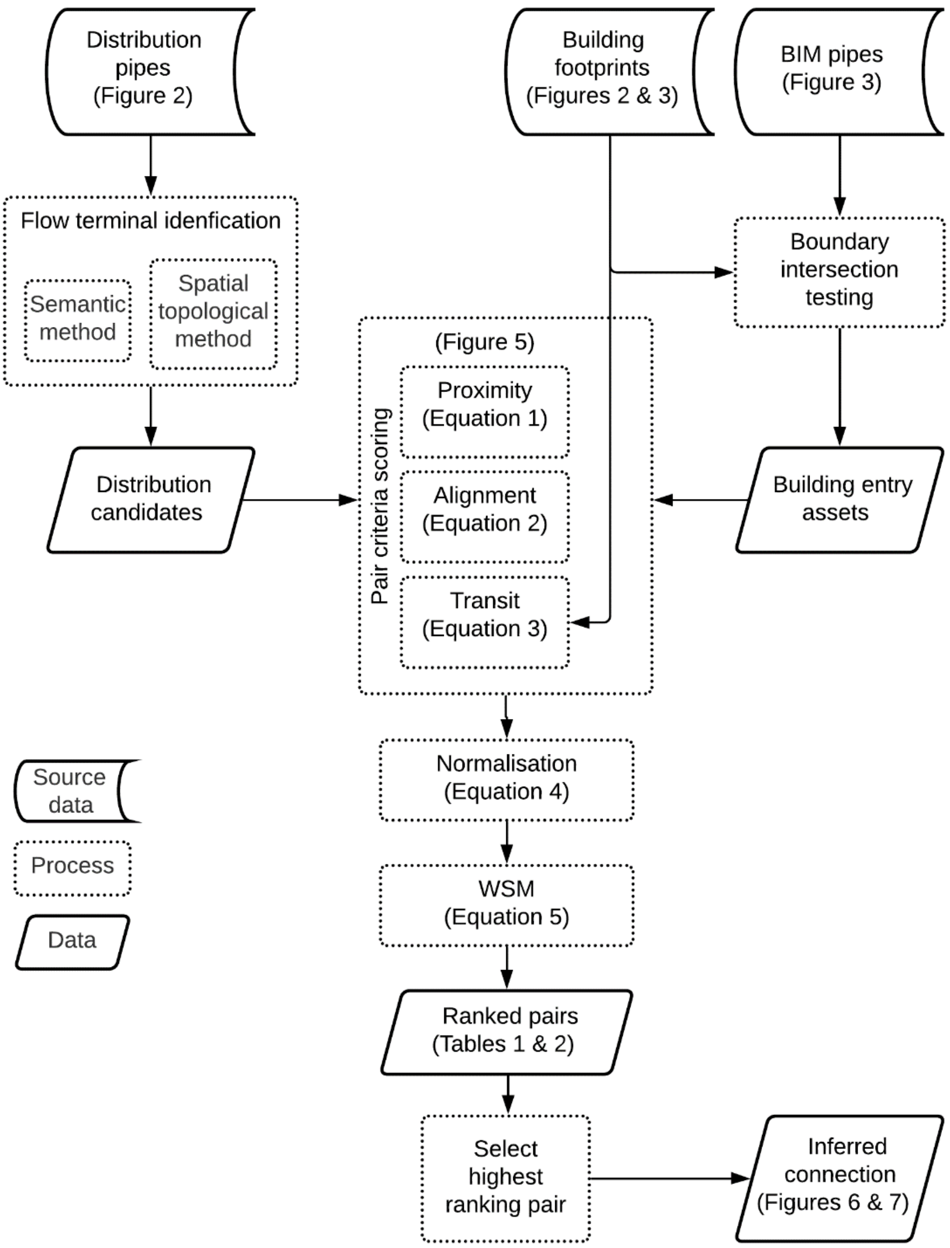

Figure 4.2.5-1 - Flow chart of the heuristic inference algorithm, which makes use of three data sources. The building footprints are used for both identification of pipes that breach the building

envelope and calculation of the transit criterion – a measure of how much building footprint would be crossed by a direct connection. Scores are calculated for each possible pairing, normalised and summed with the highest scoring candidate deemed the most plausible.

A synthetic asset is created between this asset's end point and the outermost point of the asset that forms the entry point to the building; this synthetic asset transits a direct line (shortest path) between the assets. Figure 4.2.5-2 depicts the three criteria, which are now formalised mathematically.

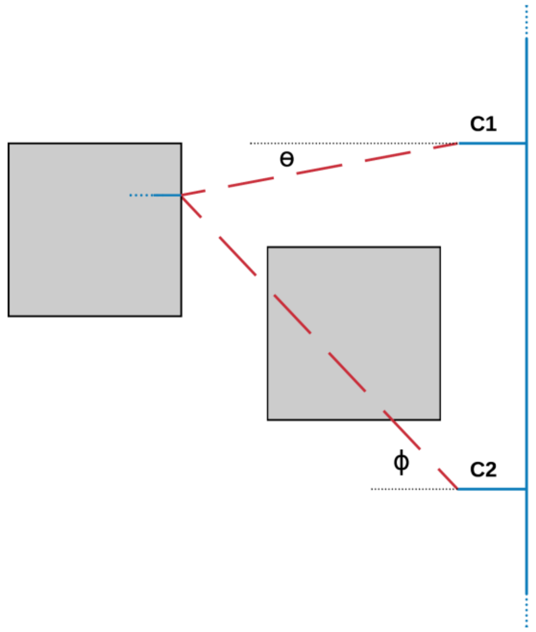

Figure 4.2.5-2 – An artificial example that explains how the criteria are scored. The red dashed lines indicate inferred connections to candidate distribution assets C1 and C2, with the best option determined by three criteria: *proximity*, *alignment* and *transit*. C1 is a more likely candidate than C2 because the inferred connection is shorter, it points more directly towards the building entry point (ϴ < ϕ) and it does not intersect any building footprints (grey areas).

The first criterion (Equation 1) concerns the proximity of the BIM asset to the candidate distribution asset. Each candidate is scored by the inverse of the distance (in 3D) between its flow end point and the start point of the subject BIM asset:

$$P_i = \frac{1}{d_i^{D \to B}} \qquad \textit{Equation 1}$$

where P_i is the *proximity* score assigned to the flow end point of distribution candidate asset i and $d_i^{D \to B}$ is the scalar distance from the flow end point of candidate i to the flow start point of the BIM envelope asset.

The second criterion (Equation 2) concerns the alignment of the final segment of the candidate distribution asset with the flow start point of the BIM entry asset; the alignment of this pipe segment with a vector that joins it to the building entry. As with the *proximity* score, this calculation is performed in all three spatial dimensions. The *alignment* score is calculated using the dot product of the two vectors:

$$A_i = \hat{\boldsymbol{v}}_i^D \cdot \hat{\boldsymbol{v}}_i^{D \to B} \qquad \textit{Equation 2}$$

where A_i is the *alignment* score assigned to distribution candidate asset i, $\hat{\boldsymbol{v}}_i^D$ is the unit vector of the pipe segment at the flow end of asset i and $\hat{\boldsymbol{v}}_i^{D \to B}$ is the unit vector from the flow end point of candidate i to the flow start point of the BIM envelope asset.

The third criterion (Equation 3) concerns the total amount of building footprint (summing over all footprints) that would be transited by a pipe (in 2D) running directly between the candidate distribution asset and the BIM asset. Given the uncertainty in position of the distribution assets and recognition that direct transits are a simplification, a binary condition that mandates zero intersection is not used; instead, plausibility is again calculated probabilistically – less intersection is deemed more likely.

The use of an inversion operation to model the lower plausibility of high values would result in extreme and misrepresentative values in cases where there is very little intersection and division by zero where there is none; instead, the value is given a negative sign to reverse order but maintain magnitude. For any building flow entry point, the following equation describes how *transit* scores are assigned to each distribution candidate:

$$T_i^{\cap} = -\sum_{j=1}^{n} \boldsymbol{v}_i^{D \to B} \cap \boldsymbol{A}_j^{FP} \qquad \textit{Equation 3}$$

where $T_i^{\cap}$ is the *alignment* score assigned to distribution candidate asset i, $\boldsymbol{v}_i^{D \to B}$ is already defined (above) for the *alignment* criterion, ∩ indicates quantitative continuous intersection (not qualitative binary), $\boldsymbol{A}_j^{FP}$ is the footprint area of building j, and n is the number of building footprints in the calculation.

The criteria scores are then normalised using the following min-max feature scaling equation (Equation 4):

$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \qquad \textit{Equation 4}$$

where X takes on criteria P, A and T. As a means of multi-criterion decision analysis, a weighted sum model (WSM) (Equation 5) was then applied to the normalised criteria for each candidate:

$$S_i^{WSM} = \sum_{c=1}^{n} w_c a_{ic} \qquad \textit{Equation 5}$$

where S_i^{WSM} is the WSM score assigned to candidate asset i, c is a criterion (*proximity*, *alignment* and *transit*), n is the number of criteria (three for this study), w_c is the weighting factor for each criterion c, and a_{ic} is the normalised score for candidate i and criterion c. In the absence of strong evidence or reasoning for prioritising any criterion over another, the weighting factor w_c is set to unity (a value of 1) for each of the three criteria, such that they are attributed equal importance. The distribution candidate asset with the highest value of S is then selected as the most plausible candidate for pairing with the BIM entry point asset.

4.2.6 Technical implementation

The geospatial data, which comprises NWL's geodatabase and OS Topography and VectorMap Local layers, are processed within Safe Software's Feature Manipulation Engine (FME) and output GML files, ensuring a consistent format for subsequent analysis. Two FME workspaces were developed for filtering out redundant attribute data, subset the datasets spatially, drape them over the LiDAR digital terrain model (DTM), offset the NWL assets beneath the surface and convert the required feature data into the target GML files. The attribute data required from pipe asset features are their identifiers, the BNG coordinates of the line segments that represent their geospatial positions and the network function. For the OS building footprint data, the DescriptiveGroup and DescriptiveTerm attributes are used to select all features that are building outlines and, for the Topography layer, the PhysicalPresence attribute is used to distinguish between ground contact footprints and overhangs for the Helix buildings (see section 4.2.2). The vertices of the polygons representing these building outlines are retained for spatial location. Represented in GML, these data form the geospatial datasets used for the analysis.

The IFC data for the Helix buildings are processed entirely within a Python (v3.7) application, making use of several modules that were developed for this project and several external libraries. The IfcOpenShell library is used to parse each of the IFC models and store the IFC elements in memory. The script subsets the IFC element data according to the requirements of the specific analyses; for example, IfcFlowSegment assets belonging to a main water system are selected for the selection process detailed in 4.2.4. Functions were developed that trace the element relationship structure in the IFC files necessary for obtaining individual object geometries, and for converting between the Cartesian CRSs local to these objects and the OSGB36 geospatial CRS needed for integration with the OS footprint and NWL asset data. The DTM model used for the geospatial data is also applied to correction of the elevation of the base coordinate (within the IfcSite element) of the IFC model for B3. Python code was developed to carry out the translation, rotation and scaling operation used to bring the IFC data into the OSGB36 CRS within the same processing pipeline. This coding of the conversion process, rather than use of a software package, enabled more insight into data quality and characteristics through debugging within the PyCharm integrated development environment (IDE) and visual inspection of geospatial maps output at intermediate steps.

Within the Python project, the NWL (geospatial) and IFC assets are represented in custom 'geobim' objects for which custom object and attributes classes were developed. The classes are designed to represent the object types attributes required for the analyses: GeoAsset, IfcAsset and SyntheticAsset (for inferred connections). The UML for these classes and their attributes for these classes are detailed in Figure 4.2.6-1. A module was developed for generation of these custom objects once the source data had been parsed. Given the size of the IFC models (262-851 Mb each), these in-memory Python objects for the IfcAssets are then dumped into Pickle files (.PKL) for reruns of analyses and visualisations, avoiding unnecessary repeats of time-consuming IFC parsing (processing time for the geospatial data was not inhibitive).

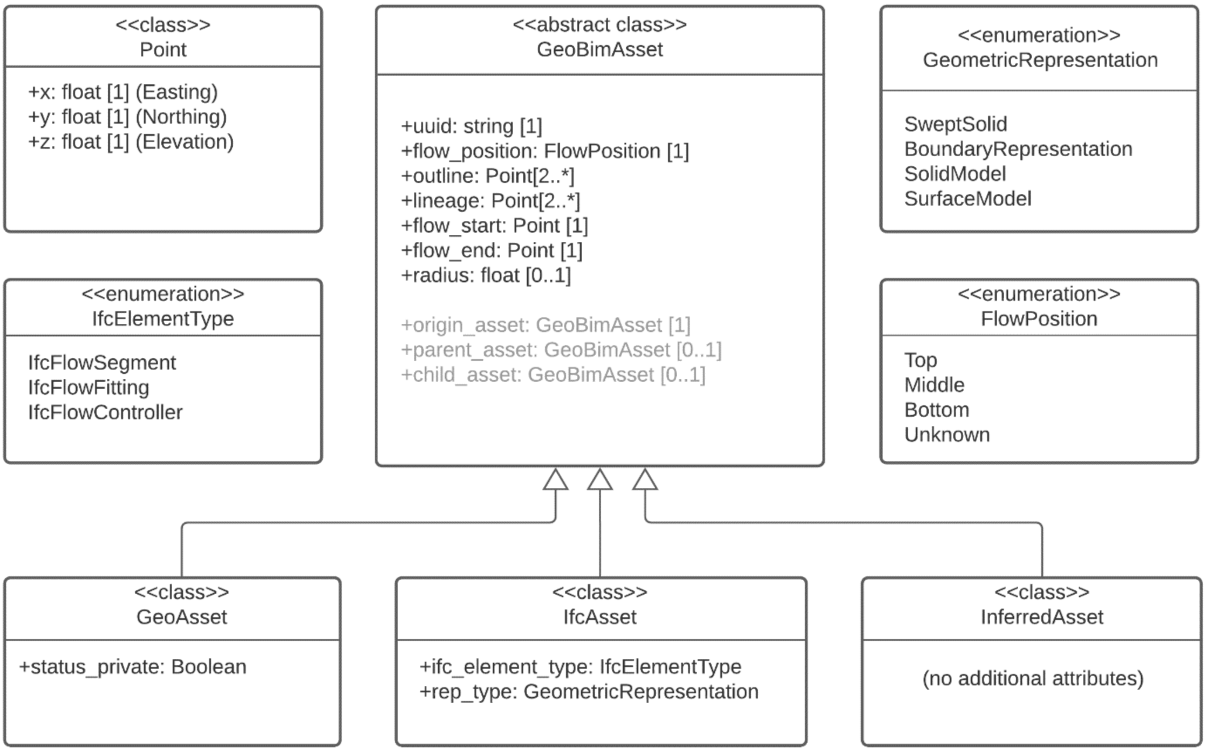

Figure 4.2.6-1 – A Universal Modelling Language (UML) representation of the data model developed for integration of the WDN with the BIM water networks. Some details of the model that are present in the project code (see impact statement on page IV for details on where to find the code) are omitted due to redundancy and some class/attribute names have been changed for clarity. The grey-out attributes of the abstract class GeoBimAsset are included to show how within-BIM flow chains were recorded as part of a smaller study, which attempted to trace flows through the internal pipework of the study subject buildings and relate them to sensor data.

The UML of Figure 4.2.6-1 shows a non-redundant representation of the data model developed for the WDN-BIM integration. The flow_start and flow_end attributes were derived from spatial topological relations (for example, the end of a terminal pipe not touching another pipe is the flow end) used for calculating precise values for *proximity*, *alignment* and *transit*; the values depend on which end point is considered for an inferred connection. These start and end points, and the flow_position attribute, do not account for the reversal of flow direction that can occur in some parts of WDNs; their purpose is to identify flow direction with respect to connectivity down from geospatial scale to BIM scale so WDN assets in which flow direction could change are not considered. Lineages and outlines were computed from the geometries in the source data. The GeometricRepresentation enumeration values were used to determine whether the outline or lineage of the assets should be used for geospatial mapping: for SweptSolid, the lineages were used, with the radius attributes used to inform display thickness; for the others, outlines were used – for simplicity, minimum rectangular bounding boxes were used in place of the detailed boundaries. The IfcElementType was used to filter for the elements for those that should be considered for intersection with the building footprints (only the IfcFlowSegment elements were considered). The status_private attribute for the GeoAsset class represents whether the NET_FUNC attribute was set to PRIVATE.

There are many ways a building or facility can be modelled in IFC, such that the specific detail omitted in this filtering or abstraction process may vary between the implementations for different case studies. In this case study, the IFC elements are embedded in complex, hierarchical spatial and thematic structures, and many objects are represented with very detailed boundaries that primarily served visualisation purposes. Most of this complexity was discarded (and hence lost) during the abstraction of these elements to instances of the IfcAsset class shown in Figure 4.2.6-1.

The comprehension, processing and visualisation of IFC geometries required significantly more programming work than for the geospatial data. SweptSolid elements (used for pipes) were translated to lines, making use of the objects' local orientation and length; Brep, IfcFacetedBrep, SolidModel, SurfaceModel, and IfcFaceBasedSurfaceModel elements were resolvable to geometric boundary representations from which bounding boxes (that contained the spatial limits of the surfaces) were generated for visualisation.

Several Python modules were written for generating maps of the source data (section 4.2.7), carrying out the analyses (including the algorithm of section 4.2.5) and producing the results (section 4.2.2) of this chapter, making use of multiple external libraries including Numpy, Matplotlib, Shapely and Geopy. Other Python libraries, including NetworkX, are used for the subsequent network analysis detailed in section 4.3. The project is structured into separate modules for statistical, spatial and network analysis functions (along with others for spatial mapping and graph network visualisation). Git was used for version-control and the repositories are described in the Impact Statement of this thesis.

4.2.7 Resultant transboundary topology

The process described early in section 4.2.4 returns all distribution network connection candidates and one building entry asset for mains water for each of B1 and B2. The georeferencing of B3 is adjusted (by a rotation of the project vector, as described earlier) until an entry asset is identified, bringing the assets into alignment with the corresponding footprint. The output of this first process is a one-to-many relationship between BIM entry points (one per building) and candidate distribution assets. The algorithm described in section 4.2.5 is then applied, identifying which of the many distribution candidates is the most plausible for pairing with the single building entry points. The candidate selection algorithm is applied to the source data with candidates identified firstly using spatial topology and then using semantics, as described in section 4.2.4. The results are presented in Figure 4.2.7-1 & Table 4.2.7-1, and Figure 4.2.7-2 & Table 4.2.7-2.

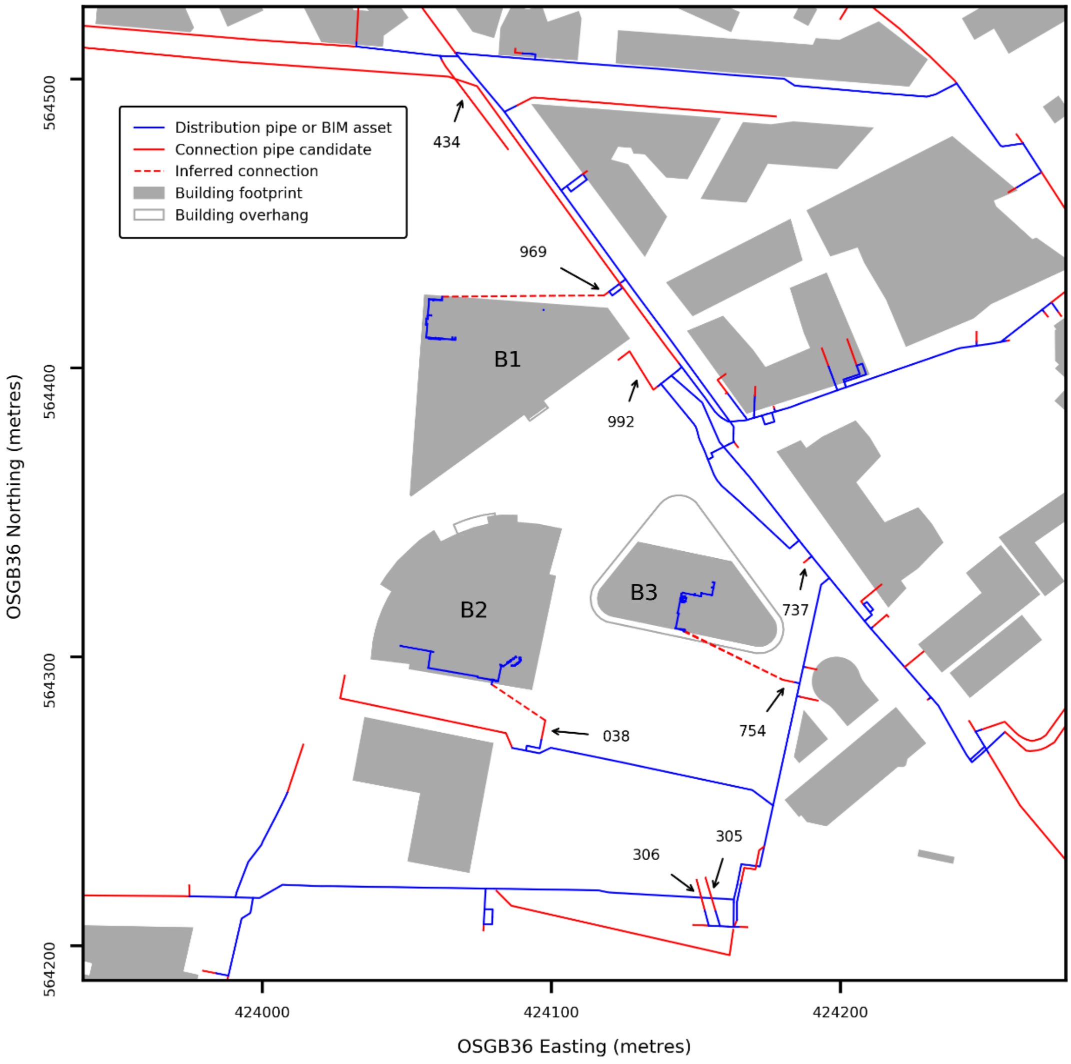

Figure 4.2.7-1 - Results of applying the algorithm described in section 4.2.5, with distribution assets identified by their spatial topology. The dashed red lines indicate the most plausible connections. The three-digit figures identify the distribution assets that are candidates for connection to buildings B1, B2 and B3 (see Table 4.2.7-1 for scores). The ID values for the pipe assets (e.g. 969, 038 etc.) are the final three digits of the IDs used in the NWL dataset. Contains OS data © Crown copyright and database rights 2020 Ordnance Survey (100025252).

Building	Asset ID	WSM	WSM diff	Proximity	Alignment	Transit
B1	969	2.901		0.998	0.902	1.000
	434	2.742	0.159	1.000	0.742	1.000
	992	2.353	0.389	0.854	0.770	0.729
B2	038	2.692		1.000	0.692	1.000
	306	2.127	0.565	0.200	0.927	1.000
	305	2.124	0.003	0.196	0.928	1.000
B3	754	2.982		1.000	0.983	0.999
	737	2.643	0.340	0.792	0.993	0.858
	038	2.506	0.137	0.656	0.851	1.000

Table 4.2.7-1 - Plausibility scores for the top three distribution candidates for each building when identifying candidates by their spatial topology. The WSM diff is the difference in weighted sum model score for that candidate and the one just above in rank.

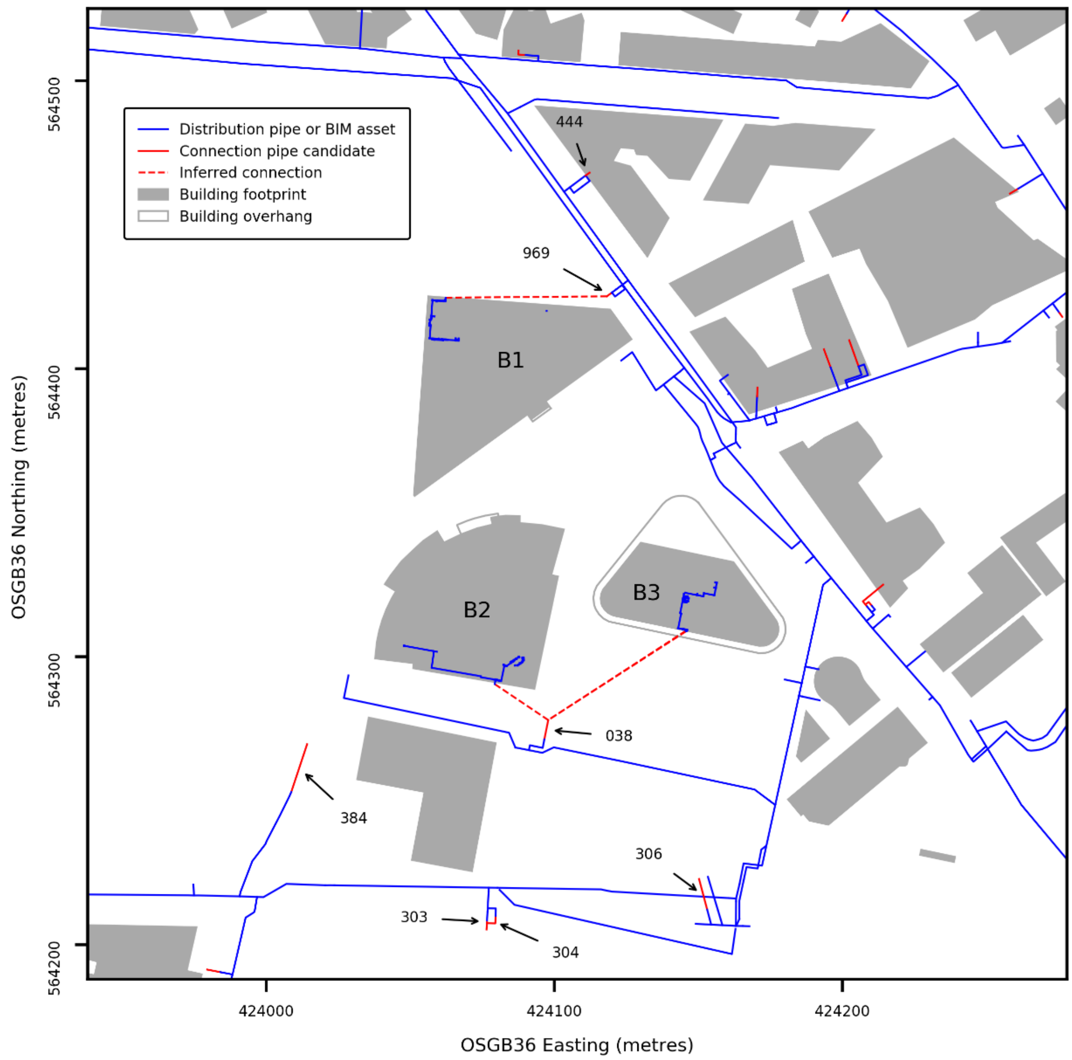

Figure 4.2.7-2 - Results of applying the algorithm of section 4.2.5 with distribution assets identified by semantics. The three-digit figures identify the distribution assets that are candidates for connection to buildings B1, B2 and B3 (see Table 4.2.7-2 for scores). The ID values for the pipe assets (e.g. 969, 038 etc.) are the final three digits of the IDs used in the NWL dataset. Contains OS data © Crown copyright and database rights 2020 Ordnance Survey (100025252).

Building	Asset ID	WSM	WSM diff	Proximity	Alignment	Transit
B1	969	2.903		1.000	0.903	1.000
	384	2.095	0.808	0.257	1.000	0.839
	444	1.819	0.276	0.824	0.003	0.992
B2	038	2.692		1.000	0.692	1.000
	304	2.233	0.459	0.243	1.000	0.990
	303	2.184	0.049	0.238	1.000	0.947
B3	038	2.859		1.000	0.860	1.000
	306	2.606	0.253	0.606	1.000	1.000
	304	2.332	0.274	0.397	0.934	1.000

Table 4.2.7-2 - Plausibility scores for the top three distribution candidates for each building when identifying candidates by semantics. The WSM diff is the difference in weighted sum model score for that candidate and the one just above in rank.

The real-world connectivity of each of the buildings to the WDN – the ground truth – is known as far as is evidenced in detailed building services schematics, from site inspections and (for B1 only) the study of flows between the WDN and the USB (detailed in section 4.2.3). A section of the diagram showing connection of B3 on its southeast side is shown in Figure 4.2.7-3.

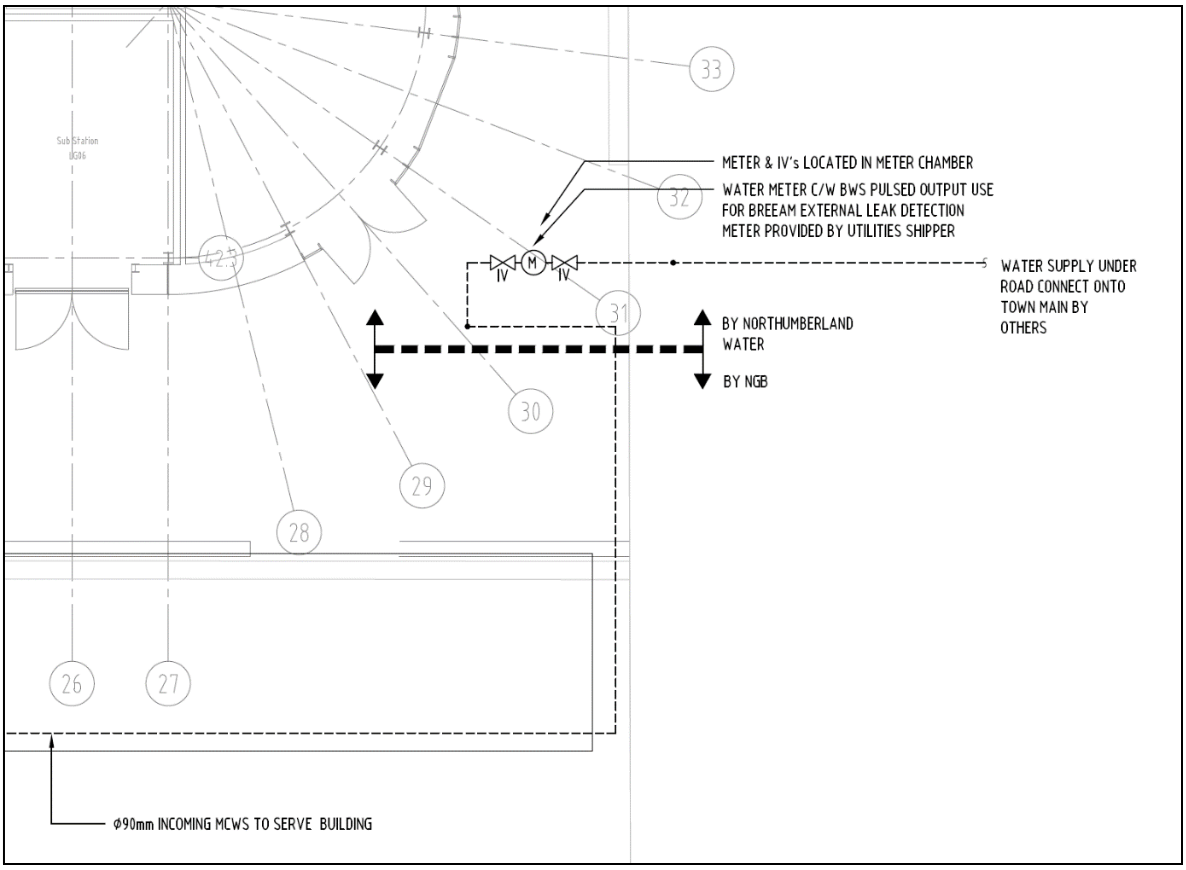

Figure 4.2.7-3 - Section of a diagram showing that B3 connects on its southeast side. This and other similar diagrams were used in combination with site visits to ground-truth the connection points of each building to the WDN. The dashed black line running eastwards from the meter chamber is approximately at the location of asset 754, shown connected to B3 in Figure 4.2.7-1.

In both sets of results, the distribution asset identified as most plausible is 969 for B1 and 038 for B2; these selections are consistent with the ground truth. An indication of confidence in this selection is provided by the weighted sum model difference (WSM diff in Table 4.2.7-1 and Table 4.2.7-2). For B1, the selection confidence is substantially higher when identifying candidates using semantics rather than spatial topology (WSM diff increase from 0.159 to 0.808); for B2, the WSM diff changes from 0.565 to 0.459, indicating that the use of semantics does not increase the confidence in selection of asset 038. However, only the relatively poor alignment of asset 434 with the vector connecting its end point to the entry of B1 (*alignment* criterion) prevents its selection as the top candidate identified through spatial topology. The results in Table 4.2.7-1 for B1 show that the Euclidean distance (proximity criterion), if used alone, would return asset 434 as the most plausible; the results also show that the transit of asset 992, which would implausibly run through the footprint for B1, ensures it is a relatively weak candidate.

For B3, the selection differs between the two results: when identifying candidates using spatial topology, asset 754 is considered a candidate and the algorithm identifies it as the most plausible (Figure 4.2.7-1); however, when using semantics, asset 754 is eliminated as a candidate and the most plausible candidate for connection to B3 is asset 038, the same as for B2 (Figure 4.2.7-2). The ground truth is that B2 connects to (or at least through) asset 754, as calculated for candidates identified using spatial topology; furthermore, this selection is more confident than the selection of asset 038 from the candidates that are identified semantically, as indicated by the high difference in WSM between the top two candidates (0.340 compared with 0.253).

In order to assess the impact of the uncertainty in position of the distribution assets on the results, the sensitivity of the algorithm to this uncertainty is measured using a Monte Carlo method. The positions of the distribution assets are randomly and repeatedly adjusted across a uniform distribution within the maximum Euclidean distance error of 6 m; the result of 1000 iterations on the top three candidates for B1 (when spatial topology is used for candidate identification) is shown in Figure 4.2.7-4 as an example.

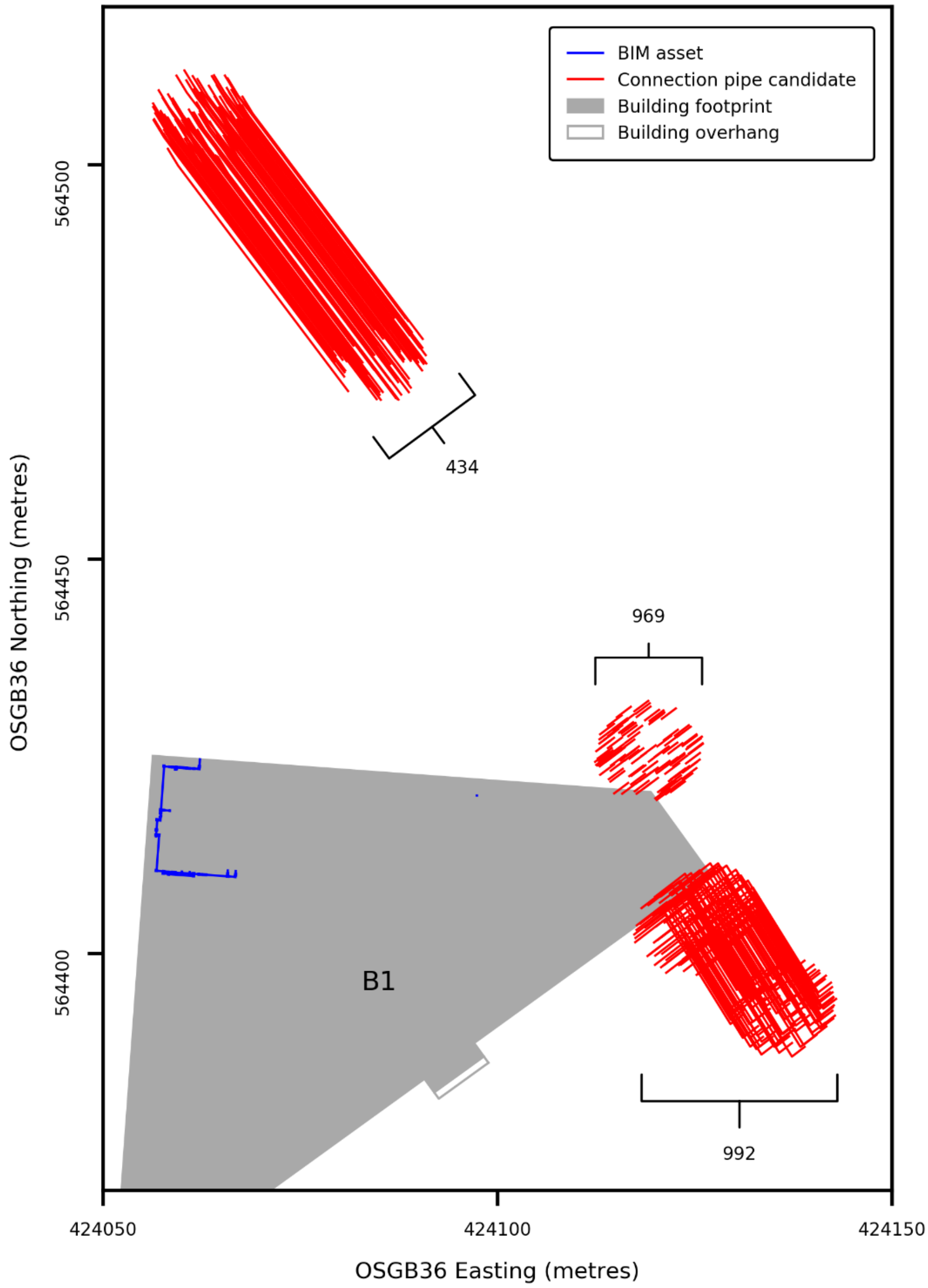

Figure 4.2.7-4. Positions of the top three distribution candidates (identified using spatial topology – see section 4.2.4) for B1 after repeated randomisation of their positions within the measured Euclidean error bounds of ±6 m. This sensitivity analysis was applied to all candidate assets for all three buildings (results in Table 4.2.7-3). Contains OS data © Crown copyright and database rights 2020 Ordnance Survey (100025252).

The algorithm is then applied to the source data for each of these repetitions, generating statistics (Table 4.2.7-3) that represent potential alternative outcomes under the uncertainty conditions. For B2 and B3, the candidates selections remain unchanged, regardless of whether the candidates are selected using spatial topology or semantics. However, depending on the candidate selection method, the most plausible asset returned by the algorithm was not consistent for B1 – these differences are shown in Table 4.2.7-3; when spatial topology was used to identify candidates, asset 434 was returned as the most plausible for connection to B1 for approximately 10% of the iterations (demonstrating sensitivity to uncertainty in position) but when semantics are used, the selection is always asset 038 (demonstrating insensitivity).

	Spatial-topological identification				Semantic identification			
Building	Asset ID	WSM mean	WSM range	Total selections	Asset ID	WSM mean	WSM range	Total selections
	969	2.847	0.435	905	969	2.878	0.503	1000
B1	434	2.701	0.276	95	384	1.972	0.145	0
	992	2.141	0.455	0	444	1.715	0.226	0
	038	2.691	0.244	1000	038	2.688	0.262	1000
B2	306	2.091	0.135	0	304	2.177	0.285	0
	305	2.092	0.137	0	303	2.121	0.403	0
	754	2.990	0.059	1000	038	2.865	0.099	1000
B3	737	2.530	0.398	0	306	2.527	0.252	0
	038	2.520	0.398	0	304	2.233	0.146	0

Table 4.2.7-3. Results of a sensitivity analysis in which a Monte Carlo method was used to randomly and uniformly vary the position of the distribution assets within their maximum measured error in position (6 m) over 1000 iterations.

The algorithm can correctly infer connections between all three buildings and the WDN without the use of semantics; the use of semantics yields higher confidence on correct selection for two of the buildings but an incorrect selection for the third. The results of the sensitivity analysis (Table 4.2.7-3) indicate that, for the study dataset, the algorithm is sensitive to existing known uncertainties in position of the distribution assets when candidates are identified using spatial topology but insensitive when semantics are used accurately and completely in the source data (with reference to right-hand side of Table 4.2.7-3, note that B3 is assigned to asset 038 incorrectly due to incomplete semantics).

4.2.8 Discussion of the heuristic inference method

The method has applicability to the use case described in section 2.2.3 (water network partitioning) by providing a heuristic method of inferring the point of connection of finer-scale BIM networks to the WDN. In the context of dynamically configuration WDNs, this allows the simulation of down-stream impact of real-time topology reconfigurations. The characteristics of or element types within the BIM MEP modelling may indicate some criticality to continuity of supply to the facility represented. Using the example given in 2.2.3, the MEP model of a hospital may contain many objects representing life-support appliances. Along with a digital representation of where this facility connects to the distribution network, this information indicates can be used to predict the consequences of cutting off or reducing supply (or pressure) to a node in the WDN at which the hospital connects.

It should be noted that the study considers only three buildings in a single urban area. Generalisations of the analyses and assessments of the results of this study remain tentative until the method (or similar method) is tested on datasets containing a larger number of BIM models of different facility types (residential, for example) and for different geographic regions; this scaling will test whether the method has been unintentionally over-fitted to the case study source data, help to evaluate the assumptions that underpin the method, support unequal criteria weightings and identify other criteria for the MCDA. For example, similarity in diameter between pipes in candidate pairs could be used to increase selection confidence; trees and other vegetation (or urban features) that penetrate the ground beyond a threshold depth could be modelled as obstructions to safe and stable pipework transit, given that the growth and movement of tree roots is known to cause damage underground pipes (Cameron, 2001; Pritchard, Hallett and Farewell, 2013); and linear features such as paths, streets and channels could be used to determine a transit that is more likely than a shortest path (it is also clear from the datasets used in this study that pipes are placed along more circuitous routes that run alongside structures). For the transit criterion, some open areas enclosed by footprints may be implausible transit zones; convex hulls of footprint vertices could be used to address this. 4D BIM adds scheduling data as the fourth dimension and might enable a modelling of the time-dependency of physical obstruction. However, any refinements should be balanced against potential loss of generalisation and hence transferability to other types of utility infrastructure.

The application of the method to a larger dataset brings into question the scalability of the inference algorithm with respect to speed of computation. This is important to the application of the method to realistic scenarios (see section 2.4.8) – such as entire urban areas – for which larger datasets must analysed. With the existing code implementation and hardware used, the inference algorithm takes approximately eight seconds to complete on the study dataset. For most existing building stock, BIM models do not yet exist and new BIM models (for existing or new facilities) could be processed as they become available rather than in bulk, such that this processing time would not likely be inhibitive. If processing speed did need to be increased, the distance from a building entry point within which WDN assets should be considered for candidacy (200 metres was used for this study) and the number of footprints that could be intersected by an inferred asset could both be reduced. Furthermore, the algorithm could be parallelised by distributing the computation by BIM-WDN pair (and more powerful computers could be used), such that the estimated potential computational demand would not be prohibitively high even for bulk processing of realistic urban areas. However, more constraining might be the number of BIM models in need of a manually correction similar to that carried out for B3.

The transferability of the method to other utility types should be evaluated by applying it to wastewater, electricity and gas network infrastructure. This evaluation could use the same study area as an experimental control but, as previously discussed, would also need to encompass other geographic regions and facility types to yield results that can be interpreted with more confidence. With sufficient BIM data for the building stock, the method could be applied to network optimisation over a much larger area. Any evaluation on other utilities should consider whether the existing MCDA criteria and underpinning assumptions are appropriate for these other utility types.

This study shows that it is possible to integrate networks with a heuristic inference method but also that the process is technically demanding and far from automatable with existing datasets. As described in section 4.2.4, the spatial topology of the assets needed to be used to identify candidates for connection to the building, making use of spatial inference to overcome the inconsistent use of attribution that could otherwise have identified candidates with more confidence. Although these challenges might suggest that the cost outweighs the benefit at present, if digital representations could be standardised such that they better facility trans-boundary connections, the integration could be sufficiently automated to make the process worthwhile. In order to better quantity any likely benefit, there is a need to identify other use cases for the ability to represent the connectivity of finer-scale internal building water networks with their surrounding WDNs.

4.3 Application to water network layout planning

4.3.1 Introduction

Chapter 2, section 2.2.4 outlined a configuration planning use case in which the layouts of internal building and distribution networks can be optimised in the context of each other if their digital representations can be integrated. Supply reliability can be increased by ensuring that a critical facility connects to a node in the network that offers higher routing redundancy and this section examines how this can be used as a factor in WDN-BIM layout planning. The research approach of section 2.6 specified the need for influence from data producers and application domain experts; for this study, discussions with a Continuous Improvement Lead at Northumbrian Water were used to guide and verify the relevance of the research to real-world WDN management scenarios. Conversations emphasised the importance of 'marginal gains' in reliability of water supply to consumers, particularly given the large fines incurred by provides even in the event of small increases to outage durations.

Robustness, redundancy, resilience and reliability are closely related but different concepts. A robust system may be defined as one that is less sensitive to disturbances (Homayounfar *et al.*, 2018) or more tolerant of errors and failures (Yazdani and Jeffrey, 2011). Redundancy is residual system capacity (Žiha, 2000) or the existence of alternative supply paths (Yazdani and Jeffrey, 2011) and is a key component of system resilience (Matthews, 2016). Resilience is defined in section 2.2.3 as the capacity of a system to resist, absorb, withstand and rapidly recover from exceptional conditions. Both robustness and redundancy (and hence resilience) are related to system reliability (Žiha, 2000). It follows that the topological configuration of a network influences system performance with respect to all of these qualities. In the context of WDNs, a network with high routing redundancy can suffer more interruptions (such as pipe bursts) without supply failure because there are more alternative paths to consumers that bypass the assets causing the interruption. Giudicianni *et al.* (2018) articulate this as follows: "The complex and meshed structure of WDNs allows the system to recover from failures, exploiting the topological redundancy provided by closed loops, so that the flow could reach a given node through different paths.". Although other factors influence system resilience and reliability (such as pipe construction material and age, environmental conditions and reservoir levels), it is favourable for a building's main water network to be configured such that it may be connected more easily to a point in the local WDN that offers the highest redundancy; conversely, it is favourable for a WDN to be configured in such a way that a topologically favourable connection point is close to the entry point to a building.

Water distribution networks are spatially constrained by their geography (Boccaletti *et al.*, 2006), can be considered complex and modelled as graphs (Giudicianni *et al.*, 2018), and graph theoretical approaches can be applied to them in search of robust network topologies (Agathokleous, Christodoulou and Christodoulou, 2017; Torres *et al.*, 2017; Giustolisi, Ridolfi and Simone, 2019), which are an important factor in network resilience. The following section considers how graph metrics can be used to measure the increase in WDN robustness that can be achieved by optimising the topological configuration of a set of buildings, again using the Helix site as the case study.

4.3.2 Example network layout

Figure 4.3.2-1 presents a fictitious, simplistic WDN of one District Metering Area (DMA) that feeds one building; this example layout supports an intuitive understanding of the influence of WDN layouts on BIM MEP design. The challenge is to determine which connection option from the WDN to the building (option 1 or option 2) offers the highest routing redundancy and hence which internal building layout is preferable.

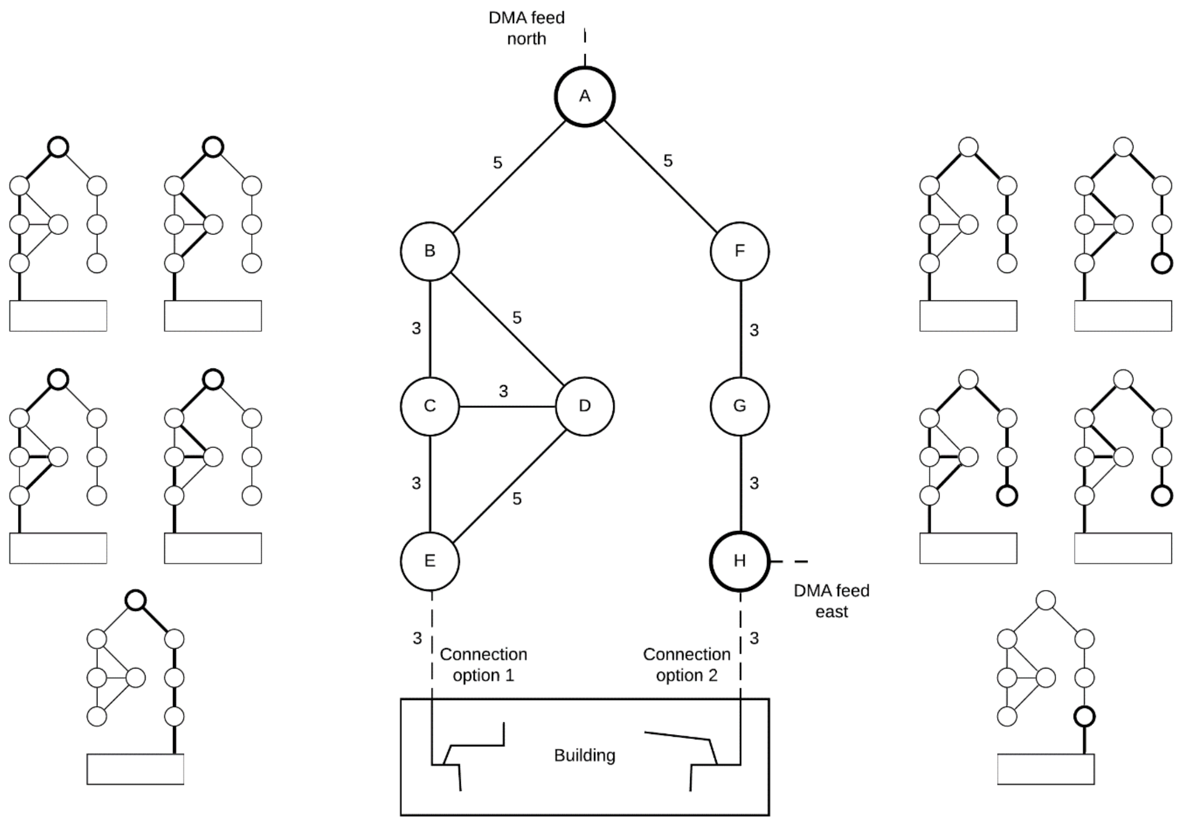

Figure 4.3.2-1 – Centre: a fictitious, simplistic WDN that is used to demonstrate how a network layout may influence the configuration of a BIM MEP model (or vice versa). The integer values next to the links indicate approximate Euclidean distances (arbitrary units). The building is treated as a single node without consideration of the internal building network (which is shown for illustrative but not analytical purposes). Left: flow paths when the DMA is fed from the north; Right: flow paths when the DMA is fed from the east.

If the DMA of Figure 4.3.2-1 is fed from the north via node A, from a routing redundancy perspective, it is intuitive that option 1 is topologically favourable to option 2 (the routing options are shown on the left of Figure 4.3.2-1): the multiple paths through C and D offer alternative paths to the building in the event of a pipe failure in the double-triangle region formed by nodes B, C, D and E. Alongside each link is shown an approximate Euclidean length (arbitrary units). Note that the route A-B-C-E is of identical value to that of A-F-G-H, such that the additional paths that use D ensure that connection option 1 necessarily offers more redundancy (for example, if pipe B-C bursts, water can transit via B-D). However, if the DMA is fed from the east via node H, although there remains only one path via connection option 2, this path is shorter than all paths via option 1, presenting higher vulnerability to supply; is this situation, option 2 is clearly favourable.

For more complex WDNs, it is unlikely that such a visual inspection will provide intuition for determining an optimal layout and an automated method is required. Various statistical and spectral graph theoretical techniques can be used to measure the characteristics of WDNs.

4.3.3 Robustness metrics

Several statistical metrics have been used for measuring topological redundancy and robustness: average node degree, link density, clustering coefficient, meshedness coefficient, average path length and central-point dominance (Yazdani and Jeffrey, 2010; Di Nardo, Di Natale, Giudicianni, Musmarra, *et al.*, 2017; Jung, Lee and Kim, 2019). The literature is not consistent on whether each of these is a measure of redundancy or robustness; however, although not identical, the two concepts are closely related: given that the availability of alternative supply contributes to topological redundancy and that robustness encompasses a system's error tolerance (Yazdani and Jeffrey, 2011) – which may include pipe failures – the topologically redundancy of a network contributes to its robustness. Spectral techniques, which use matrix representation of graph networks, can be used in combination with these statistical metrics in real-world scenarios and offer alternative metrics of redundancy and robustness. Torres *et al.* (2017) discovered particularly strong correlations between spectral metrics and WDN performance indicators. Spectral gap (SpecG) (Estrada, 2006) and algebraic connectivity (AlgC) (Fiedler, 1973) have been found to be the most representative spectral measures of topological robustness (Yazdani and Jeffrey, 2011; di Nardo *et al.*, 2018; Giudicianni *et al.*, 2018), with AlgC found to be most applicable to WDNs (Giudicianni *et al.*, 2018). Of the statistical metrics, Yazdani and Jeffrey (2012) describe the use of the meshedness coefficient as one of a simplified three-metric subset (along with AlgC and SpegG) that can be used for measurement of a network's robustness and redundancy.

When representing the DMA shown in Figure 4.3.2-1 as a graph network (with the building as a single node that connects to either E or H, and the DMA feeds not included as links), most of the statistical metrics (including meshedness) are unaffected by the choice of connection because the metrics depend on total counts of nodes and links (Di Nardo, Di Natale, Giudicianni, Musmarra, *et al.*, 2017) (this assertion also applies to real-world WDNs). The spectral metrics, however, do vary by connection choice. Also of importance is that all of the above metrics yield a value that is independent of whether the DMA is fed from the north or east (via node A or H) because none of the metrics accounts for the flows through the network. However, for dynamically configurable WDN topologies, potential routings extend beyond the limits of existing DMA boundaries and assets beyond the DMA entry nodes should be accounted for in robustness and resilience calculations. In this context, finding optimal points for connections of buildings should consider a wider spatial scope of the network that could be physically connected to the buildings.

Although it is "somewhat unrealistic to use a single metric to characterize network structures or capture a vast amount of information on different aspects of network robustness and vulnerability" (Yazdani and Jeffrey, 2012), given the above reasoning, AlgC is applied to the WDN surrounding and supplying the Helix site in order to examine whether this spectral metric can be used in support of layout optimisation.

4.3.4 Application of algebraic connectivity to the Helix site

Although it is known from conversations with employees of NWL that the WDN surrounding the Helix site is not dynamically configurable, it is now studied in a hypothetical circumstance in which flows are not inhibited by fixed states of boundary valves – that parts of the WDN outside of the existing DMA may be connected to the Helix site. The WDN is represented as a graph network with pipes as links and their intersections as nodes. Robustness metrics can then be calculated for the topological configurations that result from different options for connecting the buildings of the Helix site to this WDN.

Construction of the graph network involves exploiting the spatial representation of the pipe assets to derive a spatial topology from which connectivity is inferred. Although hydraulic models that include explicit asset connectivity for the network exist (available in INP format, which is native to the Epanet software (US Environmental Protection Agency, 2020)), these hydraulic datasets are missing data for multiple assets that are present in the geodatabase containing the WDN asset data (see section 4.2.2). Instead, the coincidence of asset endpoint is used to derive 'touch' relationships between pipes, which is assumed to imply functional (flow) connectivity. Those pipes that intersect but do not touch are not assumed to be connected (pipe depths vary). In-situ observation of pipe connectivity (based on location of manhole covers) around the Helix site imply that these assumptions yield a correct network topology but the assumption remains unverified for the entire dataset. The coordinates (in OSGB36) for the pipe endpoints were consistently represented in the geodatabase with a precision of 1mm and the same coordinate values were used for the coincident endpoints of different assets. In the graph network, any two pipe assets that touch at their endpoints are represented as two links that connect to a single node, which is identified by the shared coordinate. By this method, graph networks of subsets of the WDN are constructed. Each building and its connection to the WDN comprise an additional link (the service pipe) and node (the building). Algebraic connectivity (AlgC) calculations are performed on this graph network.

AlgC of a graph network is defined as the second largest eigenvalue of its Laplacian matrix (L), which is constructed by deducting the graph's adjacency matrix (A) from its diagonal matrix (D). The diagonals represent the degree of each node and the off-diagonals (of matrix A) the connections between each node (Yazdani and Jeffrey, 2010; Di Nardo, Di Natale, Giudicianni, Greco, *et al.*, 2017; Giudicianni *et al.*, 2018). A and L can be weighted[34] by known connection strength between vertices; this has been used for measuring the robustness of networks for air transport (Wei and Sun, 2011), satellites (Zheng *et al.*, 2017), UAVs (Nagarajan, 2018) and water distribution (Di Nardo, Di Natale, Giudicianni, Greco, *et al.*, 2017). Di Nardo *et al.* (2018) provide a formal description of these spectral graph theory concepts in the context of WDNs.

When constructing L for the Helix WDN, each link (pipe) is weighted by the inverse of the length of the pipe – the basis of this is the assertion that greater pipe length corresponds to more vulnerable and thus a weaker link (longer routes are less favourable). The connection options for each of the Helix building are the nearby flow terminals of the WDN; in this study, all of those within 50 metres of a building are allowed as options for connection to that building. The WDN is subset by regions encompassing 200, 300, 400 and 500 metres from the centre of the Helix site (see Figure 4.3.4-1), the main connected components of which are used for calculations. The increasing radius for the remits represents a larger spatial scope of the WDN that is considered to be connected to the buildings. The Helix buildings are included as nodes and their connections as links within the main components. AlgC calculations are performed for all 144 possible configurations of the three Helix buildings to the WDN for each of the bounding regions (totalling 720 calculations). The configurations with the smallest and largest AlgC values are compared for each spatial remit. The highest-AC configurations for each of the network configuration is shown in Figure 4.3.4-2 and the increase in AC from the least to most favourable configurations is shown in Table 4.3.4-1.

OS MasterMap Topography building footprint data are used in place of BIM models with the points of entry of water to the buildings left as a design parameter. For display purposes and

[34] This weighting operation also involves a scaling of the diagonals such that the sum of each row in the matrix is zero.

under the assumption that minimum pipe lengths are preferable, the BIM-WDN connections (for any particular configuration) take shortest-path transits to the perimeters of the buildings.

The technical implementation of these calculations is performed using a single Python script that uses the NetworkX library and linear algebra modules of Numpy. Much of the Python code developed for the technical implementation described in section 4.2.6 was reused. Additional functions were written for computing points on the peripheries of the building footprints that were nearest to the connection candidates. A script was developed that iterates over and runs graph analyses for all the possible BIM-WDN water network configuration options.

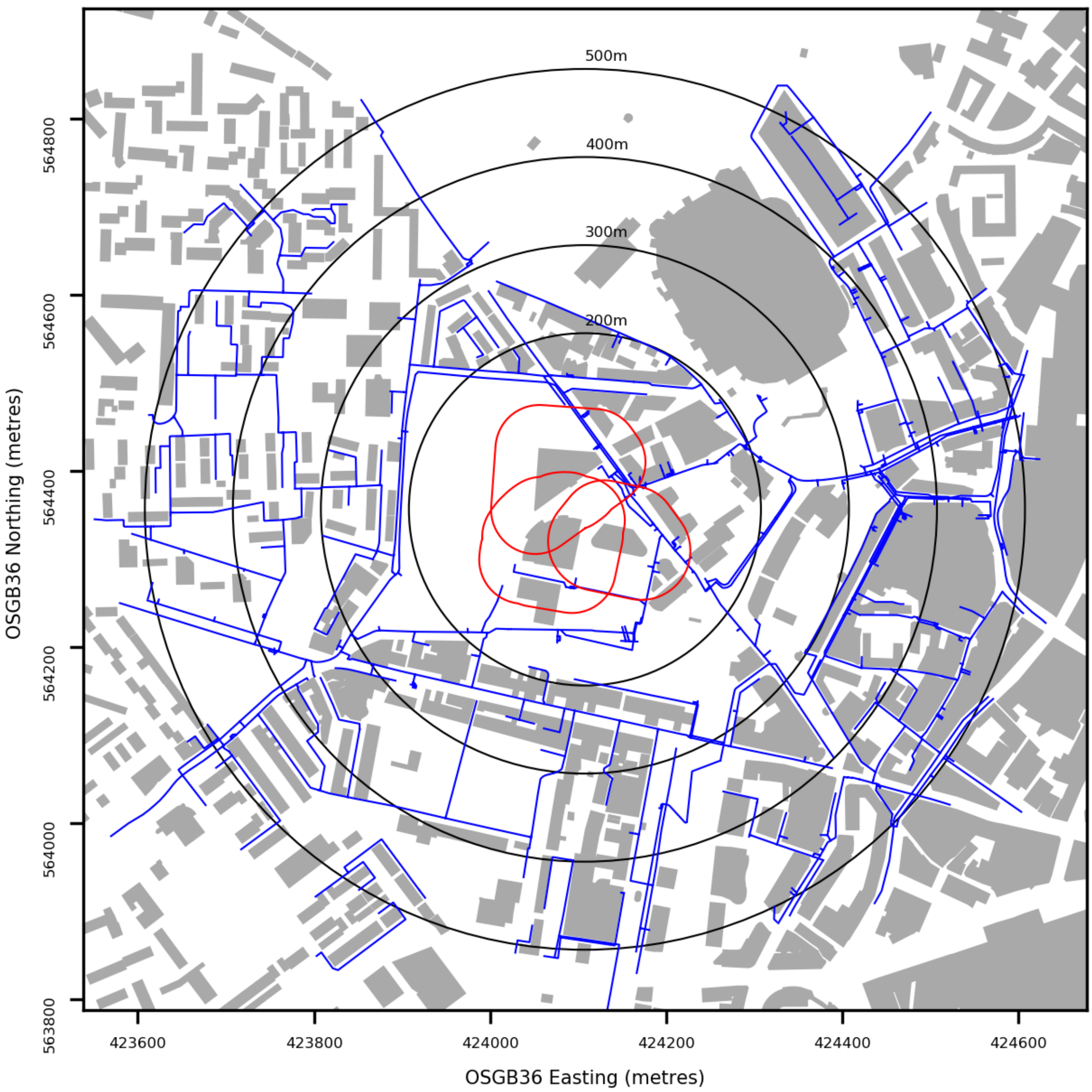

Figure 4.3.4-1 - Spatial remits (black circles) used to subset the WDN (blue lines). A graph network is formed using the spatial subsets of assets, the main components of which are used in algebraic connectivity calculations (isolated assets are disregarded). The three Helix buildings are shown in the centre with 50-metre buffers (red) used for selecting possible connections for configuration options. The other buildings in this area of Newcastle upon Tyne (also shown in grey) are not used in the calculations of this study.

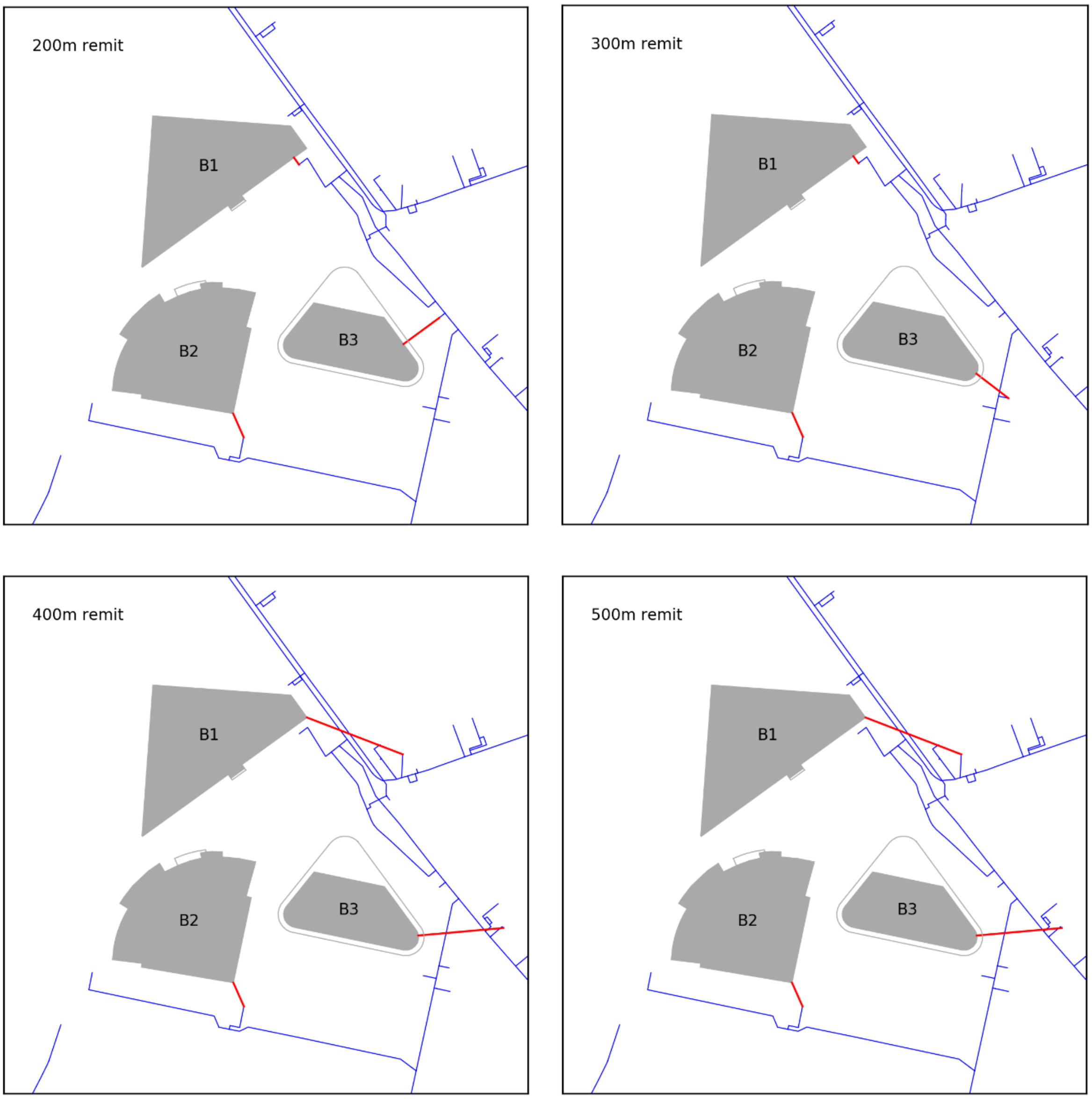

Figure 4.3.4-2 – The Helix-WDN configurations with highest algebraic connectivity (AlgC) for the main component of the graph network inside the spatial remits of Figure 4.3.4-1. The 'optimal' configuration does not change for remits above a 400-metre radius (tested up to a radius of 600 metres).

Buffer	200 m	300 m	400 m	500 m
AlgC % variation	2.39	0.761	0.127	0.101

Table 4.3.4-1 - Variation between the configuration options with highest and lowest algebraic connectivity (AlgC). The data show diminishing improvement on robustness of a WDN of increasing size given optimisation of the layout of the Helix site, which is expected given the diminishing proportion of the network that is altered by the reconfiguration.

The layouts of Figure 4.3.4-2 represent local topology that is favourable to the network robustness as a whole and the data in Table 4.3.4-1 indicate a very small increase in robustness of a WDN given changes to one local configuration (that of the three Helix buildings). The most favourable layout depends strongly on which portions of the broader WDN are involved in the calculation (those portions that fall within the respective spatial remits). Within the 200-metre remit, much of the WDN to the west of Helix is excluded; at 300 metres, the western section becomes connected and the change in B3's connection reflects this; at and beyond 400 metres, connection to the denser eastern portion of the network results in more favourable connection of B1 and B2 to WDN terminals to the east. With reference to the layouts shown in Figure 4.3.4-2, it is important to note that no account has been taken for the engineering cost or feasibility of the connections with respect to distance or the crossing of other pipes.

Table 4.3.4-1 indicates that an increase in the WDN remit decreases the whole-network benefit derived from improving the configuration in a local area, which is expected because a decreasing *proportion* of the network is reconfigured with each increase in area. The three Helix buildings comprise approximately 10% of buildings within the 200-metre bounding box and an optimisation of their configuration contributes an increase of 2.39% in AlgC. It is expected that an inclusion of all of the other buildings and their connections would result in greater AlgC changes – such calculations are not performed for computational feasibility reasons discussed in section 4.3.5.

4.3.5 Discussion of network layout planning

The results of this network planning study demonstrate that spatial layouts of BIM MEP networks and WDNs can be used in support of graph theory-based assessments of network robustness and redundancy; these assessments could comprise one component of a larger analysis and planning system. The purpose of the study is to test the feasibility and discover the difficulties of the integration challenge. Although the objective is not to prove that this BIM-WDN approach in isolation would offer a solution to improving water network topology, a discussion of the approach in the context of the quantitative results of section 4.3.4 can be used to guide a more effective implementation of the approach and identify opportunities for improving water network topology.

The results of this study show a relatively weak increase in AlgC in response to optimisation and this should be expected: when connecting to a WDN, buildings are added to the periphery of a network without providing additional connectivity between existing regions of the WDN. Despite the potential value of marginal gains (see section 4.3.1), the results might not be indicative of any significant topological improvement to the network as a whole. Far greater increases to robustness and redundancy would be expected from service pipes that connect otherwise disconnected clusters of the WDN via buildings. For example, if the Helix buildings were joined to each other via 'intra-site' service pipes (bridge connections), they would collectively benefit from the redundancy offered by potential supply from the WDN sections to their northeast and southwest (see Figure 4.3.4-1). Pointing forward in this section to Figure 4.3.5-1, it is also clear that by connecting the mains water network of building B1 to that of B2 or B3 such that they may service each other, both of the connected buildings would benefit from redundancy offered by DMAs A1 and A2 through this 'bridging'. The internal building network layouts determine the feasibility of such inter-building connection; conversely, the networks in each building could be configured (at an early design stage) to facilitate such servicing.

Regardless of how the configuration or connectivity of a network is optimised for increased redundancy, the use of a single metric (in the case of this study, AlgC) to measure any such gains is not realistic; "…the vulnerability of complex networks is assessed through scenarios of sporadic, common cause, and cascading failures, implemented by random component removal or targeted attacks on the hubs or the most central nodes/links, followed by measuring the operational consequences of such failures" (Yazdani and Jeffrey, 2011) and the use of graph theory metrics more generally "…may serve to complement traditional physics-based computer models of [water distribution systems] by providing inexpensive proxies on system-level performance" (Torres *et al.*, 2017). Any analysis that is expected to provide meaningful results of practical relevance will likely need to account for multiple topological metrics, dynamic behaviours and hydraulic modelling.

Computational complexity needs to be considered because it can be a limiting factor for more elaborate analyses over the larger datasets representing full urban areas. Despite only considering three building, the compute time for arriving at the single layout shown in the bottom-right of Figure 4.3.4-2 (50-metre remit) was several minutes. The practical scalability of using any spectral graph metric for BIM-WDN layout robustness calculations is of concern. WDNs may consist of thousands of assets (Perelman and Ostfeld, 2011) and the dimension n of the matrices used for their representation is equal to the number of nodes. The complexity of eigenvalue computation is $O(n^w)$ where $\sim 2.37 < w < 3$ (Williams, 2012; di Nardo *et al.*,

2018) and the number of unique WDN-BIM connection configurations is $O(b^c)$, where b is the number of buildings connecting to the WDN and c is the number of possible connection options for each building (if a fixed value). For an area containing 1000 buildings, 10,000 WDN nodes and 10 possible connection options for each building, 10^{30} eigenvalue calculations would need to be made for matrices of dimension 10^4. Without substantial parallelisation and distributed computing, it may not be feasible to iterate through all possible configurations and compute spectral metrics for realistic urban areas as a means of identifying topologically optimal configurations for the entire WDN.

A consideration that has similar consequence to the proposed bridging of DMAs through buildings is whether the WDN is dynamically configurable (for example, through opening or closing valves), how extensively it can be reconfigured and within what timeframe. This is directly relevant to use case 2.2.3. It was reasoned in section 4.3.3 that, in the context of dynamically configurable WDNs, more of the network than an existing partitioned area should be included in graph metric calculations. However, there is no evidence of widespread implementation of dynamic water network partitioning (WNP), it can't be assumed that all valves in a such a network can be dynamically controllable and hydraulics will constrain flow directions such that not all consumers are functionally connected with all parts of the network. The availability of different regions of the WDN for supply to a specific premises is constrained both by actual rather than potential connectivity (such as boundary valves being open or closed) and the flow dynamics of the system given these settings. The flow dynamics also depend on the physical characteristics of the network (such as elevation changes, pipe diameters and frictional coefficients) and pressure variability (Wright *et al.*, 2014).

Another consideration of perhaps greater importance is the usefulness of knowing local layouts that lead to increases in whole-network robustness: it is unlikely that opportunities often arise for configuration of an entire urban area and the owner of an individual new-build site would be more concerned with reliability of supply to their premises than a small increase to the robustness of the city's entire network. A method of identifying topologically favourable configurations to individual premises or local areas of a WDN is likely to be more computationally feasible and valuable in practice.

As previously indicated, the WDN around the Helix site is not dynamically reconfigurable. The DMAs boundaries are set by the state of manually adjustable valves. Under these circumstances, and disregarding the potential for manual changes to the status of valves, the most favourable Helix-WDN configuration is constrained by DMA boundaries and the location of the feed into the DMA – a situation similar to that shown Figure 4.3.2-1. The hydraulic models provided by NWL for their WDNs can be used to simulate flows. These models are represented in INP files, which are native to the Epanet 2 software (US Environmental Protection Agency, 2020) and not easily integrated with the NWL geodatabase. Due to the lack of integration of the hydraulic and geospatial data, entry points to the DMAs are identified through a manual inspection of the results of a hydraulic simulation run in Epanet 2 under normal network conditions. The DMAs for the Helix buildings and the metered flow entry points (M1 and M2) to these DMAs are shown in Figure 4.3.5-1. Also shown in Figure 4.3.5-1 is a connection configuration the three buildings. This configuration is calculated by assessing the flow redundancy offered by connection to network terminals within 50 metres of each building (the same as those shown Figure 4.3.4-1). For each connection option, shortest paths (the top 5 is arbitrarily selected) from the DMA entry meter (M1 and M2) through the DMAs (A1 and A2) and to each of the buildings (B1, B2 and B3) are traced. These paths are inverse weighted by both their length (shorter is better) and the independence of links (pipes) between paths (more sharing between paths reduces the value of the link) and then summed. The assets offering the highest measured redundancy are also those offering the shortest path; accounting for multiple shortest paths and independence of links did not affect the results.

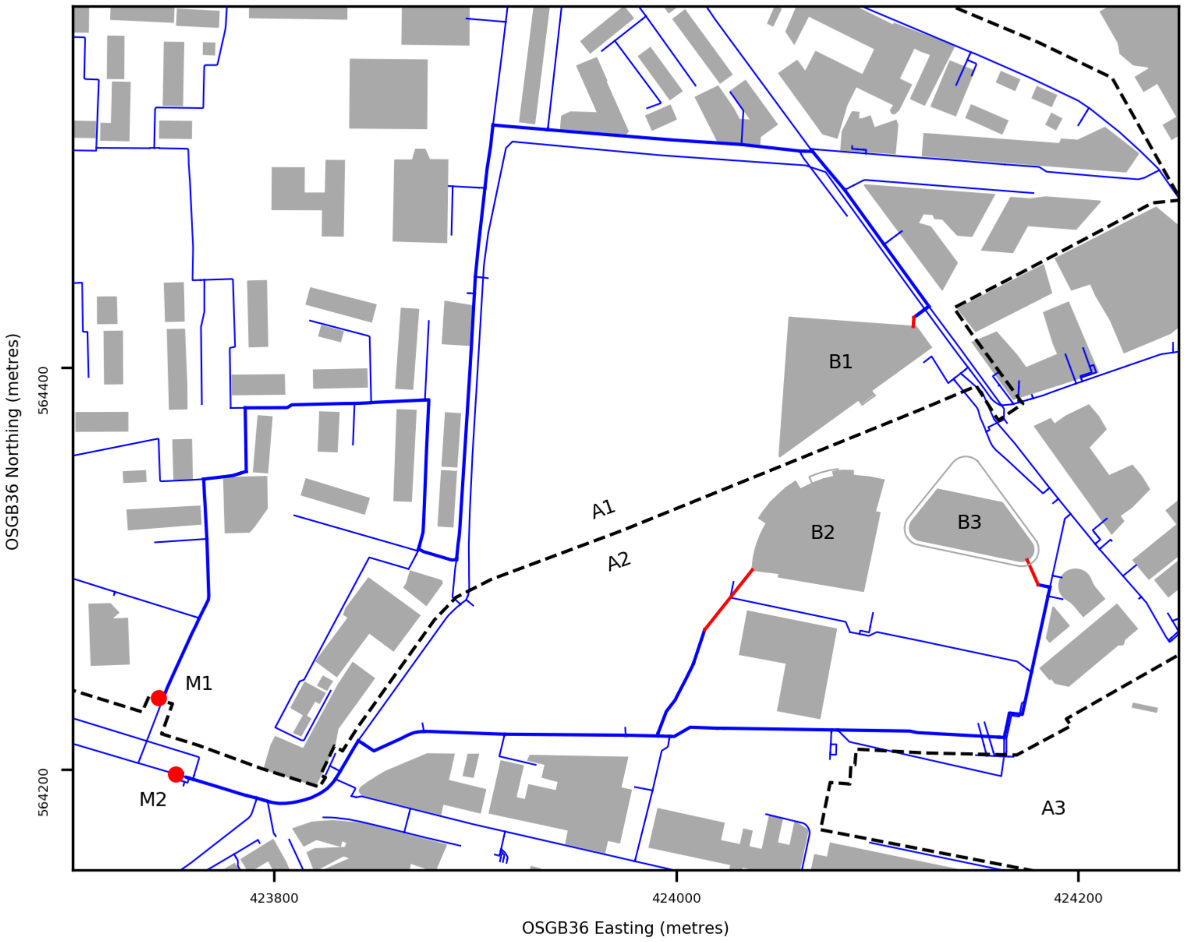

Figure 4.3.5-1 – Potable water flow paths for the Helix site, which is spread across two different District Metering Areas (DMAs). The dashed black lines delineate the DMAs (A1, A2 and A3). The red dots indicate DMA flow entry points. The red lines indicate the WDN connection points offering the most favourable connection for the buildings, based on the redundancy calculations described in the text. Water flows from M1 to B1 within A1 and from M2 to B2 and B3 within A2. Bold blue lines indicate the shortest paths between the metered entry points to these connection points.

A comparison between Figure 4.3.5-1 and Figure 4.2.7-1 (showing the actual real-world connections) shows a discrepancy only for B2. However, no allowance is made for the relative engineering cost or feasibility of this alternative connection, which is significantly longer than the real-world connection and it crosses an existing pipe. A comparison between Figure 4.3.5-1 and Figure 4.3.4-2 reveal strong disparities between favourableness of configurations depending on whether whole-graph spectral metrics or local fixed-DMA redundancy calculations are used (without consideration of physics-based factors). This demonstrates the importance of accounting for WNP and consequent limits of flow paths through a WDN that offer redundancy to consumers.

More visually discernible from Figure 4.3.4-2 and Figure 4.3.5-1 is that the choice of WDN connection strongly influences the point on the building circumference that offers the shortest path. As described in section 2.2.4, the engineering cost of any given connection option needs to be balanced with any improvements to topology. A study of Figure 4.2.7-1, Figure 4.2.7-2 and Figure 4.3.4-2 reveals that none of the existing water network entry points for the Helix buildings is situated at positions that are nearest to any of the potential WDN connection points. It is not clear that the architects or engineers have accounted for a planned connection to the distribution network. Regardless of whether topological analysis is used to influence the choice of connection point, and disregarding physical obstacles, other layouts would have enabled shorted connections to the WDN. However, it is not known whether other architectural or engineering design factors were a stronger influence on the final BIM MEP layouts. It could be that land ownership, wayleaves and an inability to lay pipes under facilities such as car parks prohibited connections that appear optimal in an analysis that accounts for only a subset of the factors that should be considered. A 'recommender' system could provide particularly high value if it identified connection options that both minimise engineering cost and maximise supply redundancy.

The need for a connected graph representation of the entire network – from supply to individual appliances – is not demonstrated in this chapter; the more generalised use cases described in section 2.2 also do not present a strong case for a full digital representation of the connectivity within each building. However, the research does demonstrate the potential benefits of connectivity at the building envelope and an aggregation of building demand beneath (topologically) this interface. Furthermore, with reference to an earlier point in this section, the finer granularity connectivity and flow directionality within buildings could be used in support of intra-site bridge connection through which buildings may offer each other additional redundancy; if the pipes in one building were to be connected to those of another, it would be necessary to know which of the assets that are close enough to be feasibly used for connection are also in an appropriate position in the flow hierarchy. There are also ‘Industry 4.0’ use cases that would similarly benefit from such detailed internal facility connectivity modelling for fault tracing a diagnosis – this is discussed more in section 5.3.5.

4.4 Discussion of the data challenges

Multiple challenges and potential pitfalls concerning missing data, inaccuracies, inconsistent units, non-standardised attribute names and non-universally unique identifiers were encountered throughout the case study.

Most of the BIM data used in this study served as an accurate and precise representation of the real-world construction. The need for manual correction of B3's orientation (section 4.2.2) demonstrates the vulnerability of BIM data to single georeferences; other studies have highlighted that georeferences are usually set to zero, default values or rough approximations (Ohori *et al.*, 2018), and that there is a "lack of control [in] the way georeferencing is stored" in IFC files (Noardo *et al.*, 2020). It would be valuable for BIM models to include further georeferencing data that enables corroboration or contradiction of project base coordinates and orientations. In the absence of this, a resurvey of at least the base coordinate for the project might be recommended for a confident exploitation of the BIM data. When BIM georeferencing is known to be inaccurate (or not present), it has been shown that accurate building footprints can be used for correction of BIM models (Diakite and Zlatanova, 2020); a convex hull of internal building BIM MEP element vertices could be leveraged similarly even in the absence of full architectural detail. The use of building footprints within the integration method of this study raises the subject of their topographic representation – the use of attributes or line features to identify the nature of polygon features (section 4.2.2) is a workaround that suggests a need for more richness of attribution. Another consideration is the assumption that entry point assets for BIM MEP networks should intersect a 2D footprint boundary because this disregards the possibility that the envelope is breached underground and inside the 2D boundary, which would require analysis in 3D, or that the MEP modelling could extend beyond the building envelope. An alternative approach would be to identify BIM element types that correspond to building entry points, such as mains water stop taps, but the reliability of this dependency should be considered.

Another issue encountered in this case study was that the network topology within the study's IFC data does not always preserve flow direction, such that the directionality of element relationships could not be used identify building connection points at the top of the flow chain. It was demonstrated that semantics can be used to identify flow endpoints but the weakness of this is a lack of consistency and standardisation, an issue that was mitigated in this study using spatial-topological analysis. This discontinuity of internal dataset connectivity limits the extent of exploitation. The scarce coverage of water usage through the building's water network downstream of the basement storage tanks (see section 4.2.3) inhibits flow simulation – the usage of water is not monitored for most appliances and there is no aggregation for building zones or floors (as for the electricity data described in section 3.3.4). Furthermore, likely inaccuracies of the water meters (again, see section 4.2.3) limit the use cases for which the flow data could be used. For example, a reduction in total measured consumption (or unconsumed flow) further down the flow hierarchy might imply leakage but only if the metering errors are sufficiently small; the converse situation (of a higher reading further down the hierarchy) could be used to identify faulty or inaccurate meters.

As mentioned in section 4.3.4, in addition to the geospatial dataset, a hydraulic model was also provided for the case study. When constructing Figure 4.3.5-1, the DMA entry points and flow paths to the buildings were derived from a combination of graph analysis for the geospatial data and manual inspection of the outputs of simulations of the hydraulic model in Epanet 2 (the geospatial and hydraulic source datasets are not integrated). Some of the data from the hydraulic model of the study WDN were integrated with the geospatial data (pipe diameters, for example) but it was discovered that many of the assets recorded in the geospatial dataset were missing from the hydraulic model[35]. Furthermore, roughness or frictional coefficients were not recorded in either dataset; an average value for the Darcy friction factor for the pipes was estimated as 0.033 using the Darcy-Weisbach equation and data provided by NWL for pipe headloss, diameter, flow rate and length but it was decided that physics-based hydraulic modelling was out of the scope of this thesis. An additional but more minor complication was the inconsistent use of units (both millimetres and inches were used for pipe diameters), use of free text for attributes values, and the use of propriety and potentially opaque or ambiguous attribute names ('NET_FUNC' is an example – see section 4.2.4). Finally, the WDN assets are identified uniquely within the datasets (with values such as WTY-MN1234567) but with no guarantee of global uniqueness. Although surmountable, these data challenges are important because they limit the extent to which the proposed integration methods can be automated.

A further consideration is the importance of maintaining the ability to infer connectivity from the locations and geometries of assets beyond an initial data integration. If there is likely to be a need to infer connectivity at a later date following (for example) changes to the layout of

[35] These geospatial and hydraulic datasets are being integrated within a single software system by Northumbrian Water during the writing of this thesis.

internal building networking or the surrounding urban infrastructure, location data should be retained and suitably managed throughout the lifecycle of a building.

4.5 Summary of Chapter 4

Water network partitioning and network configuration planning are use cases for the integration of BIM and GIS utility network topologies. The water networking in and around the Helix site was used as a case study for the development a heuristic method for inference of connectivity between internal building network and their surrounding WDNs. It was shown that this inference of connectivity can be used to reveal likely connections between buildings and WDNs, supporting analysis of the downstream impacts of dynamic network partitioning. Given that the study was applied to a small dataset for only three buildings, interpretation of the results remains tentative and the method should be tested over larger areas, in other regions and on facilities of other types. The method would be improved by accounting for other underground obstacles in the heuristics. The computational complexity of the implementing algorithm is not expected to be a prohibitive factor in applications to realistic urban areas.

The Helix site's water networking was used again for a study of the effect of different building-distribution network configurations on changes to the spectral metric of algebraic connectivity, demonstrating that spatial layouts of BIM MEP networks and WDNs can be used in support of graph theory-based assessments of network robustness and redundancy for different layout options. It is suggested that physics-based, hydraulic modelling and real-time usage data should be used to identify the flow-connected portions of the WDN to be included in graph theory calculations. It is also suggested that more of the building stock should be used in such calculations but that the usefulness of knowing whole-network improvements is questionable given that individual development projects are more likely to be concerned with local improvements. The computational complexity of spectral graph theory metric calculations is identified a potential limitation on practical implementation, contrary to the expected feasibility of scaling of the inference algorithm.

The incorrect georeferencing of BIM data, non-preservation of flow direction within BIM MEP networks, lack of integration and incomplete overlap of geospatial and hydraulic models, absence of some hydraulic attribution, inconsistency in spatial units and the use of proprietary attribute names are identified as data-related challenges that inhibited integration for the target use cases.

Chapter 5 Discussion

5.1 Introduction

The aim of this research was to devise and prototype integration methods that elicit an understanding of how existing datasets and data standards can be leveraged for construction of utility network topologies across the building envelope (see section 1.4). The research questions were as follows:

1. How can existing datasets be leveraged to construct digital representations of utility network topologies across the building envelope and how does this support the priority use cases?
2. How can existing data standards be modified to support the priority use cases where existing data cannot be leveraged to sufficient effect?

These research questions were written to account for gaps identified in existing research, which are described section 2.5 of the literature review; they include appropriate information retention and abstraction, selection of technologies in support of suitable data structuring and how data standards can be modified where they currently inhibit integration. More specifically, the literature review identified an opportunity for more consideration of enabling the simulations of multi-scale resource flows and analyses of favourable network topology. Opportunistic exploitation of conceptual commonalities, common identification of real-world objects and consistent, accurate terrestrial positioning of utility assets were identified as technical aspects that merit further research.

The following research objectives were set to answer the research questions:

1. Develop a set of priority use cases to understand the need for integrated digital representation of multi-scale utility network topologies.
2. Carry out a review of the domains of 3D Urban GIS and BIM with a focus on the representation of utility networks at and around the scale of the building envelope.
3. Examine the relevant data standards in these two domains and identify the key disparities that present a challenge to the integration of utility network datasets.
4. Analyse and critique existing methods of urban data integration, focussing on weaknesses in their applicability to utility network topology, and identify research gaps.
5. Address the research gaps through the exploration of case studies.

6. Design and prototype methods that satisfy the general requirements the use cases through addressing specific requirements of case studies.
7. Discuss the findings from the case studies in the context of the use cases and data standards.
8. Highlight areas of future research.

The first four objectives were addressed by the literature review of Chapter 2. Three key use cases that elicit an understanding of the need for integrated representations of utility networks across the building envelope were chosen and detailed. The 3D GIS and BIM domains were reviewed, followed by a study of the commonalities and disparities between CityGML and IFC, which were used as proxies for data standards in the built environment domain. The handling of spatial data uncertainties within these domains was also reviewed. Existing methods of GIS-BIM integration were then reviewed with a focus on the context of utility network topologies (section 2.4; objective 4). The methods were categorised by the techniques they implement; they were then critiqued, and the remaining challenges interpreted, in the context of disparities between data standards. The chapter then identified research gaps and devised a research approach for addressing these research gaps.

The fifth and sixth objective were addressed by Chapter 3 and Chapter 4. The research approach included the use of two case studies centred around Newcastle's Helix site. Chapter 3 focussed on the use of graph database, message broker, and web technologies for addressing the first use case of demand-side management through visualisation of electricity flow, exploiting a level of conceptual commonality between the underpinning data standards as a basis for this integration. Chapter 4 addressed some weaknesses identified in the approach of Chapter 3 and concentrated on the use of the spatial representation of utility assets for the use cases of water network partitioning and configuration planning. For both case studies, the methods that were devised and prototyped enabled the semi-automatic integration of inter-dataset topologies. Aspects of the workflows that required manual intervention highlighted opportunities for better utilisation of the data standards by data producers, enhancements to the data standards and approaches to data standardisation, and further development or refinement of the integration methods.

The seventh and eighth objectives are now addressed in this chapter through a discussion of the research findings from the case studies of Chapter 3 and Chapter 4 in the context of the research questions and gaps. This chapter draws on the research of previous chapters to answer the research questions. The dataset and standards are considered in detail in section 5.2 and then future challenges and opportunities in section 5.3.

Section 5.2.1 considers availability, accessibility and coverage of datasets; section 5.2.2 the lack of explicit connectivity between utility networks featuring in the datasets, whether this is an issue of data standardisation and how the data standards can be modified to enable expression of such connectivity; and section 5.2.3 the exploitation of identifiers for common concepts and how spatial data can be leveraged to overcome some challenges. Section 5.2.4 discusses these spatial data, focussing on the importance of uncertainties in location. Section 5.2.5 then considers integration of flow data with representations of physical utility network infrastructure, and the relevance of real-world location for this challenge.

Section 5.3.1 discusses the avoidance of underground utility strikes as a use case for the some of the research of Chapter 4 that, although not related closely to the objectives of this thesis, is of high value. There is emphasis on the applicability to recent, high-profile projects and standardisation work. Section 5.3.3 considers how the technologies and methods of Chapter 3 and Chapter 4 could be combined into a more sophisticated capability that makes combined use of all of the research of the two chapters. Section 5.3.4 then looks at further use of graph database and message broker technologies, and how they could comprise the basis of an infrastructure network modelling and simulation component of a digital twin. Finally, section 5.3.5 discusses future research topics, including the how the research of this thesis could be extended to support the modelling of multi-infrastructure (and beyond) interdependencies and methods that could be used to address coverage gaps in real-time flow monitoring.

5.2 The datasets and data standards

5.2.1 Dataset discovery, accessibility and coverage

The data principles of findability, accessibility, interoperability and reusability (FAIR) were initially conceived to "…improve the infrastructure supporting the reuse of scholarly data." (Wilkinson *et al.*, 2016). The OGC now uses these principles in its mission statement on location data. The discoverability or findability of datasets is clearly an initial limiting factor in their exploitability. Within the case studies of this thesis, an early hurdle to integration of digital representations of the utility networks in and around the Helix site was access to existing datasets. The significance of this is more acute given that the Helix site is a new development with a focus on urban research (see section 3.2). For Chapter 3, the electricity distribution network data was available as an output of other research on heuristic derivation from road transport network data; representations of known real-world layouts were unavailable due to a combination of incomplete modelling, commercial sensitivity and security constraints. For the case study of Chapter 4, the existing relationship between Newcastle University and Northumbrian Water was important for enabling access the water distribution network (WDN) dataset. However, a non-disclosure agreement needed to be signed to address security concerns, with specific permission allowing publication of results. Given that Newcastle University is the project owner for the Helix construction project, exploitation of the BIM data representing the internal building networks was implicitly permissible and access was provided through a software portal.

The existence of detailed modelling for the Helix site is likely due to it being a sophisticated, research focussed, public sector, new-build project. The Helix site serves as a study subject for some of the research teams it accommodates (see section 3.2) such that detailed modelling was a requirement for the Helix construction project. Given government funding, modelling of the site in fully collaborative 3D BIM (as a minimum) was obligatory (Cabinet Office, 2011). It is unusual for older, existing building stock to be modelled in BIM and rare for the modelling to contain such detailed MEP data. This lack of coverage by BIM of the broader building stock presents a significant barrier to scaling out of the method developed in this research to larger urban areas, for which most construction will have taken place before the existing of BIM as a discipline. However, it is unlikely to be economically feasible to retro-model the entire building stock, so an objective should be to incrementally increase the extent and sophistication of BIM modelling of new building stock and ensure that the datasets are available and usable. However, there needs to be financial incentive for this beyond immediate project requirements, which is a recurring theme throughout this discussion.

Of most significant hindrance was the difficulty in sourcing data for the assets in the space between the building envelope and the WDN and the absence of real-world location data for these assets. Approximately one week was taken to acquire data describing these assets, using contacts within at prime contractor, subcontractors and the Estates team within the university. The CAD data for the connection pipes were mostly 2D diagrams in PDF format and without any georeferencing or representation of (internal-dataset) network topology. The absence of any absolute location data for these boundary assets prevented their automated geospatial exploitation; instead, for the purpose of network topology integration, they were useful only for manual verification of inferred connections (see Chapter 4 and section 5.3 of this chapter). Although they were presumably provided to specification, the deficiencies of these datasets rendered them obsolete for the purposes of automated processing such that the region they represent is effectively missing data. If assets that connect buildings to distribution networks are to be fully exploited, their digital representation should be subject to the same fidelity requirements as the BIM MEP models. The expansion of the normal, practical BIM modelling spatial remit to include the area between buildings and distribution networks might resolve this issue.

The UML for the GeoBimAsset class (Figure 4.2.6-1) shows the IFC element data that are needed to enable the integration demonstrated in Chapter 4. The capturing of this abstracted representation during the construction phase of projects would enable network integration analysis capabilities during the operation phase. A simplified geometry, accurate geolocation and attributes that identify flow position are important for inference of inter-dataset connectivity. While Chapter 4 showed how the loss of much detailed information does not prohibit a functional integration, the capture of simpler representation is less burdensome that the capture of high levels of detail, and the retention and proper management of such simpler data would support the network topology component of a Digital Twin.

5.2.2 Inter-dataset network connectivity

Although the locations, geometries and attributes of assets described in the previous section may be sufficient for inference of connectivity between datasets, one of the key findings of this research is that there is a lack of explicit connectivity that would allow higher confidence integration. Although it could be argued that this is due to missing coverage around the interface of the datasets (see section 5.2.1), regardless of such gaps, there are no attributes in any of the study datasets that could be used to assert connectivity directly and reliably. For example, the use of the string value 'PRIVATE' for the 'NET_FUNC' attributes on mains water pipes in the WDN data identifies pipe segments that do not belong to NWL. These were used in the study of Chapter 4 for semantic identification of flow end points and thus candidates for connection to building networks.

There is also a problem of inconsistent usage of suitable attributes even when they are available within a schema. The study of Chapter 4 found that the attribute that identifies the private status of asset in the WDN was not used consistently. It is relevant to note that, even in the event of consistent use of this asset, its value would not identify the specific building network to which it connects; the attribute can only be used to generate many-to-many relationships between

WDN pipes and buildings from which a pairing must then be selected. Exacerbating this problem is that the BIM models did not identify which pipe assets were building entry points (and thus candidates for pairing). The lack of preservation of flow direction in the network topology of the IFC data forced the use of building footprint data instead: intersection of pipe and building envelope was used to identify flow start points. The heuristic spatial inference method described in Chapter 4 was developed in response to these deficiencies. In some respects, the inference method succeeded but its reliable applicability to utility network integration more generally has not been demonstrated. It might also be reasonably assumed that, even if such attributes that allow explicit one-to-one connectivity did exist, there might be little incentive for their usage (their instantiation within datasets) unless data owners are interested in supporting capabilities beyond their responsibility.

At this point in the discussion, it is worth reiterating the understanding of 'integration': in the context of this thesis and its objectives, Chapter 2 defines integration as asserting connectivity between multiple disjoint digital representations of real-world networks that are physically and functionally connected in the real-world. Notwithstanding that this definition is not universal, a suitable reference from one dataset to another might be sufficient for a shallow level of integration. This observation relates to the technique of embedded referencing described in section 2.4.3, which depends on software interoperability for deeper integration of information contained at a referenced resource or dataset. When representations of networks are abstracted to a link-node graph structure, such a level of integration can be achieved if assets reference each other between datasets. It could be ensured that this abstraction resolves to a link-node structure in which the nodes represent concepts that are sufficiently close to all users' cognitive models – this relates to the ontological abstraction technique detailed in section 2.4.6. The result of applying aspects of both techniques is explicit connectivity between instances of common concepts and the case studies of this thesis show that such integration is functional (to some extent) for the target use cases. However, the presence of an attribute such as 'NET_FUNC' with value 'PRIVATE' within the NWL WDN dataset is only suggestive of such connectivity and the heuristic spatial inference method of section 4.2 was required to demonstrate this functionality. NWL datasets appear to be structured around only the company's needs (understandably) and an application of FAIR principles (section 5.2.1) to data production would increase breadth of reusability (although probably not security). The absence of a standardised set of appropriately abstracted concepts and the inability to assert inter-network connectivity through referencing between instances of these concepts are the primary challenges that this thesis addresses.

Arising from this situation are questions relating to the inter-network information that should be provided in the source data in support of the use cases, and whether this information needs to go beyond enabling assertions of one-one-pairings and identify other qualities or characteristics of these relationships. The Model for Underground Definition and Integration (MUDDI) conceptual model includes the IGraph interface with attributes connectsFrom and connectsTo, and using this "… networks can be related to each other as either subnetworks (containment) or subordinate networks (dependency). Networks consist of nodes and links, which in turn connect to each other." (Lieberman *et al.*, 2019). Concepts such as 'House Connection' and 'House Service Line' are both being considered for inclusion in MUDDI. The UtilityNetwork ADE has the NetworkLink element (UtilityNetwork ADE contributors, 2011; Kolbe and Kutzner, 2016) that is intended '…to connect different NetworkGraphs to each other to form a Multi-Modal (multi- utility) Network" (Becker, Nagel and Kolbe, 2010). IFC does not appear to accommodate connectivity to utility networks outside of the remit of BIM. Data models need to enable the identification of a node within a subnetwork as a candidate for future connection to another subnetwork (through inference, for example) that could be represented in a different data standard.

There is another question around responsibility for populating such connectivity attributes in the datasets if the individual data owners do not stand to benefit directly or in the short-term. Is it the responsibility of the project owner, prime contractor or subcontractor? Given the likely longer-term planning and associated financial risk, this richness of dataset attribution could be mandated by legislation with costs reimbursed by government, recognising the benefits of dataset reusability to future public sector projects (such as urban development research initiatives) – legislation might also need to guarantee free access or at least a reduction in cost for data acquisition, which might need to be subject to censorship such as aggregation or partial omission because privacy and security is a concern.

The availability of standardised attributes that identify flow subnetwork termini will likely be insufficient if a higher degree of confidence the automatic integration of subnetworks is required; in such a case, each subnetwork might need to be able to reference explicitly another. However, the producers of a dataset for one subnetwork might not know about the dataset used for representing the network to which it does or will connect in the real world (the real-world object or its digital representation might not yet exist). The challenge is to enable the assertion of such a connection through reference to a future digital representation of an existing or future real-world object.

5.2.3 Object identifiers, common concepts and abstraction

Globally unique identifiers (GUIDs) can be used to ensure unambiguous reference to real-world objects – and can be used in support of integration – but there are challenges to their effective implementation. Chapter 3 shows how the merging of nodes that represent objects that are duplicated across multiple datasets implements a joining of the subnetworks of each dataset. However, as described in section 3.4.2, this commonality needed to be introduced artificially because the individual datasets use different identifiers. Furthermore, the identifiers were not all guaranteed to be unique. The proposal for more extensive use of a Unique Property Reference Number (UPRN) system in the UK is a step in the right direction. In a letter to the Secretary of State for Housing, Communities and Government, the Director of the Geospatial Commission has recommended that "It should be mandatory for all public sector data sets, relating to properties and buildings, to include the UPRN" (Propertymark, 2021). Meter Point Administration Number (MPANs) are available for electricity supply points, Meter Point Reference Numbers (MPID) for gas (UK Power Networks, 2018) and Supply Point Identifier (SPID) for water (Everflow Water, 2019) – these relate to the property or land rather than utility asset. However, such a range of identifiers can impede efforts at common identification.

Without a standardised way of commonly and uniquely identifying objects across multiple domains, a combination of spatial and semantic or conceptual information could be used. If two datasets represent, for example, a building (the common concept) and the centroids are approximately collocated, it could be assumed that the same building is represented, which could then be corroborated by matching or similar outlines/footprints (shape comparison). However, in the context of utility network integration, this point is relevant primarily to the method developed in Chapter 3; not all network integration scenarios will be related to buildings and their envelopes. Utility network dataset may be disjoint at the boundary of other facility types or within no structure that can (or would be) conceptualised, such as an underground space with no discernible boundary.

If conceptualisation of the assets (such as pipes and cables) themselves is common, reliance on such 'hook' or proxy objects (such as a building) might be avoided. An alternative approach would be to identify utility assets commonly and uniquely instead of using the properties supplied for the assets. Within the case study of Chapter 4, identifiers were not used for integration but, had attributes such as connectTo or connectFrom (as used in the MUDDI standard – see section 5.2.2) been present for the assets and populated with *common*, universally unique identifiers (CUUIDs[36]) that reference assets in the other subnetwork, the network topologies could have been joined with confidence and without inference. This absence of not just unique identifiers but *common and unique* identifiers is perhaps one of the most significant problems in this domain of research, and for digital twins more broadly. It is not clear how this could be achieved in a practical way. What objects should be identified, where could someone discover an existing CUUID and how could they generate one for a new object or one for which a CUUID has not yet been generated? The objective might be to devise and publicise a system that generates an identifier for a real-world object that is independent of who generates it and when they do this. The basis of a CUUID could be a concatenation of object type from an agreed set of object type names and a real-world coordinate expressed in an agreed CRS, and these identifiers could be accessed and populated through a distributed ledger technology (DLT) (for the purpose of information completeness and consistency rather than mistrust).

Key to such a CUUID resource would be the agreed set of object types. The development of a shared vocabulary of common concepts for the built environment would help to address this. "The buildingSMART Data Dictionary (bSDD) is an online service that hosts classifications and their properties, allowed values, units and translations." (buildingSMART, 2020). Successful examples exist in other domains, such as some open civil data standards (Azavea, 2021) that use sets of object types. For example, the General Bikeshare Feed Specification has enjoyed "…extremely broad adoption with hundreds of public and private bikeshare programs." (GBFS contributors, 2021a) and uses a standardised set of vehicle and propulsion types (GBFS contributors, 2021b). A service or resource defining agree object types in the built environment domain more broadly would increase the consistency of object representations, supporting cross-domain software interoperability and potentially providing a basis for generation of CUUIDs.

A further issue is that that complimentary datasets cannot be relied upon to each have instances of the same asset. Chapter 4 considers how to infer connections in such a situation, making use

[36] The use of UUID within the acronym CUUID is not intended to refer specifically to the 128-bit number used to identify information in computer systems; instead, the intention is simply to describe a string value that can be used commonly across multiple domains and is unique with universal scope.

of both the location and geometry of mains water pipes in the non-overlapping BIM and WDN datasets. Beyond the quality and effective use of spatial data, the success of the method depends on filtering for IFC elements of a type that matched that of the WDN assets (water pipes) – the setting up of this filter was a manual process due to some differences in the semantics used in each dataset, a process that was onerous enough to imply a need for semantic and conceptual harmonisation. In the WDN dataset, pipes were identified as mains line features with water as the resource being implicit; in the BIM datasets, they were identified as IfcFlowSegment elements belonging to a mains water system. This is an example of unnecessary and inhibiting difference. Both subnetworks concern potable mains water passing through pipes that are part of utility systems and neither dataset would be diminished by identifying them more similarly. Although some divergence of conceptualisation in different (albeit converging) domains is inevitable and freedom of real-world conceptualisation is important for allowing solutions to be developed for evolving requirements, "…concepts could be identified for opportunistic harmonisation, on the understanding that the process is more feasible for new concepts or those that need to be reworked. This would result in an incremental harmonisation of the standards." (Gilbert *et al.*, 2020). This can be summed up as promotion of gradual convergence of conceptualisation. As with ontologies (see section 2.4.6), this needs to be close to the users' conceptualisation rather than that of data modellers. If users of various use cases across the multiple relevant domains are consulted throughout the development of a standard, it may be possible to ensure that this closeness of conceptualisation is by consensus such that the standard is implemented more widely.

For some use cases, remaining conceptual differences can be reconciled through abstraction, a concept that was reviewed in Chapter 2 (again, see section 2.4.6) in the context of existing integration methods. Abstraction of detail can remove differences, yielding the commonality that is necessary for generalisation (Ponsen, Taylor and Tuyls, 2010; Kamarudin, Ridgway and Ismail, 2016). If two previously incongruous features can be generalised to the same thing, they can be analysed and otherwise regarded in the same way and thus may be considered integrated. However, this is more relevant in use cases that involve the operations of visualisation and querying of digital environments (Gilbert *et al.,* 2020) than explicit network connectivity. An additional advantage of such abstraction is the removal of information that might otherwise prevent sharing due to privacy or security concerns (such as retention of design data about a single-occupancy office space but removal of usage statistics). However, there is a balance to be struck: abstraction to a coarser granularity may increase homogeneity at the expense of functionality; a concept with reduced detail is a valid representation of larger range of objects but that lost information might be critical to some uses. Returning to an example from section 2.4.8, reducing the classes of 'pipe' or 'cable' to the more generic 'conduit' could prevent some operations, such as the functional integration of networks that contain conduits only of relevance to a particular resource type or domain. Another example concerns the preservation of flow directionality as well as connectivity, such as for dwellings that both consume from the National Grid and feed it via micro-generation.

For the purpose of integration within the case studies of Chapter 3 and Chapter 4, there were technical challenges related to the comprehension, processing and visualisation of IFC geometries, which are described in section 4.2.6. However, the more fundamental topology-related requirements of the use cases of concern to this thesis (section 2.2) are not hindered substantially by the disparities in geometries between the BIM and geospatial domains (such as the dilemma concerning the parametric representation of surface observation, which is outlined in section 2.3.5). However, the importance of the quality of real-world location information to the outcomes of integration by spatial inference is significant enough that it is dedicated a separate section in this discussion.

5.2.4 Spatial data uncertainties

The research of Chapter 3 and Chapter 4 show that, with datasets and standards in their current form, it is not possible to automate integration of network topology using object identifiers and semantics alone. This is due to both the missing data problem outlined in section 5.2.1 and the challenges concerning object identification and conceptualisation described in section 5.2.3. The case study of Chapter 4 demonstrated a way of mitigating these limitations, showing how spatial data can be used to infer connections across unrepresented sections of water networks between the building envelope and the WDN. The consistency and accuracy of location data present in the datasets of Chapter 4 were critical to the success but also the limitations of the method, which should be discussed in detail. The handling of spatial data uncertainties by existing data standards was discussed in section 2.3.6 of the literature; the research of Chapter 4 that focusses on uncertainties in location data should be discussed in this context.

The GNSS survey described in Chapter 4 verified that OS building footprint data for the Helix site are accurate to approximately 0.15 metres but the NWL WDN data only to 6 metres. Despite this difference in accuracy, the coordinates in both datasets were expressed with millimetre precision, which is misrepresentative of the accuracies – far more so for the WDN data. This problem is particularly acute because neither dataset contains any information about location uncertainty, although OS does publish this information in a separate document (those uncertainty figures were assessed as conservative). As identified in section 2.3.6, despite existing efforts, there does not appear to be any widely accepted, standardised mechanism for representing spatial data uncertainties.

One impact of not knowing the uncertainty in location is an inability to bound the scope or reliability of any integration method that depends on location. The survey carried out for the use case of Chapter 4 demonstrated this impact through a sensitivity analysis, showing that the results are dependent on variation in position of WDN assets within the measured error bounds. Given the inconsistency between precision and accuracy described above, standardised attributes are needed that can be used to state explicitly the numerical uncertainty. The use of categories for accuracies or uncertainties (see section 2.3.6) is unlikely to be generalisable across multiple domains. This discussion point is not exclusive to geospatial datasets: there is a need for IFC and other BIM standards to support a standardised representation of geospatial (and geometric) uncertainties. This could involve a standard attribute across all domains for uncertainty in each of the three dimensions, with an agreed data type for the values. There is a need for standardisation of the representation of spatial uncertainties across all domains of relevance to the built environment. There is also the potential to fuse multiple built environment datasets to estimate positional uncertainty: closer agreement (lower collective variability) between representations could be used to infer higher accuracy.

Another consideration is why the WDN data described in section 4.2.2 were accurate only to 6 metres (irrespective of any representation of this uncertainty). Through conversations with employees and ex-employees of water companies, it is understood that much of the mapping data might have been taken from paper documents, a process that is subject to inaccuracies introduced by the original printing/drawing and then interpreting the finite line widths of hard (non-digital) copies. Given the availability of portable, high-precision GNSS equipment, accuracy in positioning of distribution assets could feasibly be increased by an order of magnitude without the use of a trained surveyor. Utility companies have already equipped maintenance personnel with GNSS trackers that enable them to update the location of assets during operations. There is potential for more automation of this process: for assets with known approximate locations and that are visited for maintenance, tracking points could be collected passively (with a tracker fitted to the person without a need for their interaction). Centroids of points that are spatially clustered around the approximate location could be used to reduce the coordinate error bounds. For assets of unknown location, clusters of points spanning a period that is known to correspond to the expected time and duration of a planned, standard maintenance operation could be used to derive a set of candidate coordinates. The level of accuracy required for such assets depends on the context; section 5.3.1 discusses this in the context of spatial inference.

The georeferencing in the BIM models used in Chapter 4 did not suffer from similar *systematic* issues concerning accuracy and precision; instead, omission of or significant error in the expression of georeferencing forced manual correction. The literature suggests that this is not unusual: Ohori *et al.* (2018) state that the attributes in IFC files that are intended for georeferencing are usually populated with values "…that are almost always set to zero, to a default or wrong location, or to a very rough approximation of the real location…", a situation that is exacerbated by "…the mismatched definitions of the positive direction for the longitude in IFC2x3 and IFC4." – a negative value represents east of the zero meridian in v2x3 but west in v4. The elements within the model are then subject to the quality of this one reference, which is a significant vulnerability to geospatial exploitation without burdensome manual inspection (the standard does not provide for internal corroboration/verification). In other critiquing, Uggla and Horemuz (2018) remark that "IFC does not support the use of object-specific map projections, nor does IFC offer any means to compensate for the difference in scale between the construction site and the map projection.". In support of more accurate georeferencing for integration with geospatial data, Uggla and Horemuz (2018) recommend the addition of attributes for the expression of object-specific map projections and a scaling factor for the (x, y)-plane, and that the use of European Petroleum Survey Group (EPSG) CRSs and their codes for these projections is sufficient for most geospatial integration use cases; GML and hence CityGML provide this capability. It is apparent that the IFC standard provides the basic constructs required for georeferencing but would be improved by the addition of some attributes for map projections and scaling, and redundancy in georeferencing that allows the user to verify georeferences without manual inspection.

There must also be sufficient incentive for data producers to georeference BIM models if that task is beyond their immediate remit or interest. Use cases that are long-term or not clearly profitable might not be factored into requirements for digital representation. The potential for increased reusability and hence onward selling of georeferenced BIM models might not be well understood; addressing this could be a matter of awareness and communication of the benefits by government departments. A similar point to those of sections 5.2.1 and 5.2.2 – concerning legislative intervention or public funding – could be made for georeferencing of BIM models; for example, the Geospatial Commission could offer funding in exchange for free usage across the geospatial and AEC sectors on publicly funded projects. This might not need to be a long-term investment: once its usefulness is evidenced, data producers might choose to carry out such data enrichment by default such that public funding becomes unnecessary.

It is worth reiterating a well-known and long-standing problem concerning the ordering of CRS axes that was encountered on occasions when processing some of the datasets used in the studies of this thesis. There is a conflict between the usual expression of Cartesian coordinates in the order [x, y] and the frequent – but, critically, not consistent – expression of geospatial coordinates as [latitude/northing, longitude/easting]. Given the usual orientation of geographic data with north up, the latter ordering is equivalent to [y, x], which confuses data processing because the coordinates are not always 'switched' in this way. It usually cannot be known without inspection which item refers to latitude/northing and which to longitude/easting. There is clearly no need for this conflict and confusion. It is possible to automatically infer the order in many circumstances: if it is known that a dataset relates to a particular land region, there are limits to the values for easting, northing, latitude and longitude that place a data set in that region and a particular interpretation of coordinate order can be excluded programmatically. For example, a coordinate of (54.973459, -1.6104412) interpreted as (longitude, latitude) would place it in the Indian Ocean but in Newcastle upon Tyne if interpreted as (latitude, longitude); if the dataset is an IFC model of a building from a UK data provider, the latter is clearly more likely to be correct.

5.2.5 Correspondence between flow data and network nodes

In the same way that the lack of common, unique identification of network assets inhibits an integration via common proxy objects (such as buildings) or direct referencing between assets represented across different data sources, data streams are generally also not assigned identifiers by which they can be directly related to sensors (or objects monitored by sensors). Using the case study of Chapter 3 as an example, flow data is assigned to nodes representing internal building spaces and consumer types within the USB. The internal floor and zone structure of the building was derived from the same stream of data that contained the usage figures. The fictitious Building X (BX) was assigned demonstrative values for visualisation purposes. The study did not demonstrate the integration of flow data from one source with a network structure derived from another. The task of automating such an integration is a significant challenge that is related to the discussion points of sections 5.2.3 and 5.2.4.

Despite much external sensor data being accompanied by geographic coordinates (James *et al.*, 2020), internal sensor data are usually not, which prevents a spatial inference based on colocation (if a data stream was allocated a coordinate that matched that of a sensor element in a georeferenced IFC model, it could be inferred that the data come from the sensor). In the case study of Chapter 4, neither the sensor data describing water consumption within the USB nor the boundary valve or DMA metering data (see section 4.2.3) were provided with locations, and identifiers could not be linked automatically to real-world assets or their representation within the BIM MEP data. Although references to MEP schedules are present in the BMS data stream for the USB, these references were not present in the USB's BIM MEP model. In the absence of such spatial information and without identifiers that can be used to relate data streams from IoT (Internet of Things) outputs directly to the digital representations of their subject real-world assets or environments, it is not possible to populate a network structure with finer granularity consumption data without some level of manual intervention. Dave *et al.* (2018) propose and demonstrate the feasibility of a system for connecting IoT devices with building information data that is based on 'Open Messaging' interfaces. Shahinmoghadam and Motamedi (2019) remark that "…the integration between IoT and BIM is still in its early stages…" and that there is a need for more development of both inference capabilities and open standards (and platforms). Both research teams state the need for standardised IFC export guidelines in support of IoT-BIM connectivity. The integration of BIM and IoT data remains a significant problem for urban data modelling and the realisation of dynamic digital twins.

5.3 Challenges and opportunities

5.3.1 Suitability and limitations of the use cases

Identification of suitable and important use cases was a challenge for this research. As outlined in section 2.2.1, the use cases were limited to those that require digital representations of utility network topology. The case studies then needed to be concrete, real-world instantiations of these use cases that could test the effectiveness of an integration of utility network topology.

The extent to which these case studies demonstrated the benefits of this integration – and the limitations of such integrated digital representations – should be discussed in more detail.

Both the case studies demonstrated how the integrated digital representations are affective at satisfying fundamental parts of the requirements of the respective use cases under specific conditions, and that the methods enabled the derivation of these representations. The electricity network case study of Chapter 3 showed that a demand-supply hierarchy could be derived using the identification of building entities under the condition that the BIM and GIS models overlapped in their representation of specific buildings at their interface and that the buildings are represented semantically in such a way that they could be identified as identical – this second condition is not met in practice (the buildings were identified too differently), so the demonstrated value is hypothetical. Furthermore, the demonstration of benefit is partial because the case study did not demonstrate any actual electricity demand-side management (the use case of section 2.2.2); instead, it showed how demand can be visualised in support of such management. The case study of Chapter 4 demonstrated more completeness in addressing the requirements of the use case: the method was not reliant on hypothetical semantic consistency and the exploitation of the digital representation for water network partitioning and configuration planning was shown analytically and quantitatively – although not in a practical application. Both case studies were reasoned to be applicable to other utility types.

However, the methods developed for both case studies were limited in the types of environment or context in which they are realistically effective and practicable. It is unlikely that the demand-side management supported by the case study of Chapter 3 would be readily applicable to individual dwellings in a residential area (compared with a university campus, for example) and both case studies are dependent on high-level planning oversight and close collaboration between AEC professionals and utility companies.

5.3.2 Heuristic spatial inference and underground utilities

The heuristic inference method demonstrated in Chapter 4 has applicability beyond the use cases described in section 2.2. The method developed in Chapter 4 uses location data for pipe assets to derive spatial topological relations, from which network topology and hence a graph representation of a WDN are inferred. The chapter also demonstrated the potential to infer inter-network topology using a heuristic algorithm that made use of the proximity, alignment and transit of assets. The discussion of section 5.2.2 clarifies that this inference was a way of overcoming a lack of connectivity between datasets describing subnetworks of the utility system. The first part of Chapter 4 studied the plausibility of inferring unknown connections, showed the impact of uncertainties on results and then discussed how other underground features could be modelled as obstructing of competing transit options. If data uncertainties for other underground objects are known and attributed to their digital representations, engineers can quantify confidence in the results of algorithms that implement more complex spatial inference that accounts for these features.

Avoidance of underground utility strikes is a use case that could benefit from such inference. The underground placement of many utility assets presents complexity, engineering challenge, operational expense and risk due to lack of knowledge of their location (Deep Dig Output, 2017; Likhari *et al.*, 2017; Geospatial Commission, 2019). Collection of accurate data (such as locations) about these assets is inhibited by the difficulty of direct access (Esekhaigbe, Kazan and Usmen, 2020). A study from 2016 by the Utility Strike Avoidance Group found that, where utility plans had been studied prior to excavation, only 48% of utility assets were recorded on the plans and, of these, 84% were recorded 'inaccurately' (Likhari *et al.*, 2017). The UK's Geospatial Commission estimates an annual cost due to accidental strikes of underground pipes and cables as £1.2 billion (Geospatial Commission, 2020a) and, in 2014, the UK Health and Safety Executive (HSE) approximated 12 deaths and 600 serious injuries per year from contact with electrical cables alone (Metje, Ahmad and Crossland, 2015). The utility strike use case is a high priority for the UK's Geospatial Commission, which launched the National Underground Asset Register (NUAR) project in 2019, an objective of which was to reduce the risk of accidental strikes of underground assets by creating "…a secure data exchange platform to provide a digital map of where assets are located…" (Geospatial Commission, 2020a) – in this case, security constraints limit implementation of FAIR principles (see section 5.2.1). In the context of the broader digital twin domain, the UK's Centre for Digital Built Britain (CDBB) has placed importance on answering questions about how models are 'aggregated' despite information gaps and how unconnected assets are managed (Hetherington and West, 2020). At an international level, the Model for Underground Data Definition and Integration (MUDDI) working group (SWG) of the OGC states underground utility strike avoidance as a key point in its use case on routine street excavations; the SWG also states the "…need to know precisely where […] utilities are located in order to properly plan building foundations and new building service connections." (Lieberman *et al.*, 2019) under its use case around planning, design and construction. In areas of unmapped utilities, assessment of the plausibility of transits can be influenced by the known location of other abandoned utility features, basements, passageways and buried foundations. Increased confidence in the inferred location of active utility assets can be used in risk assessments of utility strikes; if there is a high likelihood of a pipe running underneath a section of ground, an excavator can decide not to dig there or to take more caution. For example, the building of a relatively modest basement or extension to a dwelling in a densely built-up area carries the risk of discovering underground assets that are extremely costly to move or work around, which could prevent project continuation.

Additional research could study the effect of other underground features, geology and terrain usage on the condition and vulnerability to breakage of utility assets of known or inferred location. Water pipes, especially those that are older and made from inflexible clay, are vulnerable to breakage under shear stresses caused by differential ground displacement or subsidence. Even ground movement due to traffic loads can contribute to pipe failures (Aşchilean *et al.*, 2018). There is also a potential knock-on effect in the form of bursting-induced ground displacement (Shi, Wang and Ng, 2013). It has been shown that synthetic aperture radar interferometry can measure ground movement to sub-centimetre accuracy (Xia, 2010; Wang *et al.*, 2019). Regardless of the cause of ground movement, data from such remote sensing techniques could be coupled with WDN layouts, pipe materials types and geology data in a method that predicts bursts and recommends maintenance based on comparison against patterns in data at the time of historic bust events. Section 2.2.3 describes the large amount of water lost through leakage in England and Wales (approximately 21% of public supply) and the daily global cost of leakage (approximately $39 billion). Given these figures, and that the damage caused by bursts can be significant (Wu *et al.*, 2016), the cost of targeted pre-emptive burst and leakage intervention based on confident predictions might be a worthwhile investment (Hart and Murray, 2010).

Methods involving heuristics and inference accept a level of uncertainty but it is important to quantify the uncertainty accurately. When inferences are based on spatial data, the valid scale of this inference depends on the precision and accuracy of data representation (see section 5.2.4). The sensitivity analysis of section 4.2.7 demonstrated the impact of location uncertainty on the results of an inference algorithm. For sub-meter scale inference, location need to be represented accurately with sub-metre precision. Good knowledge of uncertainties allows data to be handled appropriately and for interpretations derived from the datasets to be correctly bounded. Use cases determine the requirements on maximum spatial uncertainty and it is expected that initiatives such as the Geospatial Commission's NUAR project will continue elicit these requirements.

5.3.3 A full integration of the technologies and methods

Although relatively mature, the suitability of the specific types of technologies deployed within this thesis to utility network integration should be discussed, along with the potential for their further exploitation within the case studies. In particular, the capability offered by a synthesis of the technologies of Chapter 3 with the methods developed in Chapter 4 should be explored.

The technologies used in Chapter 3 for electricity networks could be applied equally well the water networks studied in Chapter 4. In Chapter 3, a graph database was used to represent the dynamic state of a multi-scale electricity network that was integrated via the merging of nodes with matching identifiers, and a message broker was used for sharing the dynamic state of the network; Chapter 4 showed that spatial inference can be used for direct connection of assets, thus avoiding the need for common identification of a proxy object (a building), a process that required manual intervention in Chapter 3. Although the WDN of Chapter 4 was represented in graph form for graph theory metric calculations (section 4.3), a clear next step is to create and then connect a graph representation of the internal building networks (of the Helix buildings, for example) with that of the WDN. This would result in similar but more complex graph network than those shown for Building X in Figure 3.4.3-2 and Figure 3.4.4-1 (a small example set of consumer nodes were used for Building X). Real flow data describing water flows through the Helix building and the outside urban area can then be simulated in the graph database and communicated via a message broker, as for the integration electricity network of Chapter 3. However, there are challenges to achieving this that have not been addressed by the technical work of this thesis.

Aside from the inter-dataset connectivity issues addressed in section 5.2.2, some of the hurdles to automatically generating fully integrated, dynamic representations of entire building-distribution networks concern discontinuities in the network topologies internal to BIM models, the spatially sparse collection of flow data throughout both the distribution and internal building networks, and the lack of attribution of any flow data to nodes in the network. The IFC models for the Helix site do not preserve flow direction (see section 4.2.4) and not all of the topology is recorded explicitly – the internal building subnetworks are disconnected within the models, despite being physically connected in reality. In order to derive full end-to-end topologies, it would thus be necessary to perform, for example, a spatial inference of connectivity *within* the building models, rather than just for the building-WDN connections. Given the proximity of pipes inside buildings, the feasibility of this might depend on sub-centimetre accuracy of location of the MEP modelling. Chapter 4 demonstrated that the uncertainty in position of building footprints was approximately 0.15 metres and an assumption was made, based on visual inspection, that the MEP modelling was of a similar accuracy (see the surveying study of section 4.2.2 for more details). It is possible that more precise spatial location would be needed for within-building integration. A second concern is the sparsity of monitored network nodes: within the USB, flow data is available for the mains incomer, the inlet to potable water supply tank and the inlet to the non-drinking water tank; for the distribution network, data only exist for the DMA and property boundary valves (see section 4.2.3). It is, however, perhaps infeasible to expect truly complete monitoring of networks at all scales. Thirdly, none of the available water flow data for the Helix site are attributed to (or otherwise associated with) digital representations of assets (such as IFC elements) within the Helix site and the data streams are not located geospatially, which prevents any inference from colocation – this is an example of the more general discussion in section 5.2.5. If these hurdles could be overcome, an automatic generation of the type of integrated network representations shown in Chapter 3 is feasible for water networks and potentially other utility networks. Under these circumstances, and with more building stock modelled in BIM, there is scope for a larger-scale deployment of the technologies shown in Chapter 3 for representation of the dynamic states of utility networks for realistic urban areas.

5.3.4 Suitability of the technologies to digital twins of infrastructure systems

Several technologies were deployed for implementation of the method developed in Chapter 3, most notably the Neo4j graph database and Apache Kafka message broker. Although these two products have strengths that might distinguish them from competing technologies, the technology-related interest of this research is the capability offered by graph database and publish/subscribe message brokers more generally, which may be understood through an analysis of Neo4j and Kafka specifically.

The simplicity and flexibility of graph database schemas offer a powerful capability for integrating utility networks where network topology is represented using different data structures. In Chapter 3, despite some dissimilarity in how IFC and the UtilityNetwork ADE represent network topologies (see Figure 3.3.2-2 and Figure 3.4.3-1), network features are abstracted to a relatively simple node-link-node structure; additional information or structuring is not necessary for the use cases (see section 2.2). This is consistent with an emerging recognition of the need for appropriate or targeted information loss (Stouffs, Tauscher and Biljecki, 2018; Gilbert *et al.*, 2020) to supersede a default approach of fully lossless integration (see section 2.4.8). The comprehensibility of graph database schemas (see section 3.5) ensures that they can be adapted to be closer to a user's conceptual model, which was identified in 2.4.8 as a characteristic strength of ontologies – graph representations are often used for storing the triples of ontologies. Graph databases offer a powerful way of representing integrated, abstracted network topology derived from datasets that are underpinned by disparate data standards. Furthermore, the common foundational link-node structure across graph databases technologies supports their interoperability. For utility providers, graph databases could provide a suitable means for sharing with other utility providers or for modelling the dependencies of their system on supply from these other utilities (see section 5.3.5). For both utility companies and property developers, graph databases can enable an exploitable representation of integrated networks for the use cases of section 2.2.

Message brokers are effective and efficient for storing, processing and exposing streaming data. Chapter 3 shows how a publish/subscribe message broker can be used for management and dissemination of the dynamic state of an integrated utility network, and for incremental construction of network relationships from spatial data embedded in messages. Section 3.4.5 provides some details of the deployed Kafka technology and its scalability. Further to these points, the Kafka Streams client library enables the building of micro-services for which uses Kafka clusters for storage of input and output data. Considering such capabilities more generally, streaming data of variable provenance and fidelity could be managed gracefully within such services, accounting for incompleteness of coverage and positional uncertainties, supporting the automation of more elaborate heuristic inference (such as that discussed in section 5.3.1) and digital playbacks of network flow evolution (as suggested in section 3.5), through a combined graph database-message broker architecture. And with enough data streams of appropriately attributed and standardised messages, dependency on 'static' datasets (such as BIM models) for representation of network topologies reduces (see sections 3.3.4 and 3.4.4). If the challenges discussed in sections 5.2.1 to 5.2.5 can be addressed, this architecture could form the backbone of an infrastructure network modelling and simulation system that comprises one component of a city's digital twin. For example, such a technology stack could hold the information needed for simulating a multi-scale, multi-utility cascading failure scenario that can predict fine-scale consequences of a utility supply failure, with agent-based modelling (for example) then used to simulate the consequential human behaviour, knock-on effects to other infrastructure systems and feedback effects on the utility supply networks.

5.3.5 Extended development of the methods and other use cases

Given the increasing digital representation and real-time monitoring of urban environments, the technologies and methods explored in Chapter 3 and Chapter 4 have the potential to address broader infrastructure challenges faced by the domain of smart city modelling and digital twins.

The discussion of section 4.3.5 considered a fault tracing use case for Industry 4.0 or Fourth Industrial Revolution (4IR). The increase in interconnectivity and smart automation may demand require an ability to automatically trace through functionality paths of interdependent components and services within, for example, a sophisticated manufacturing process. For example, if one process on a manufacturing line fails, it may be necessary to identify the potential root causes digitally by identifying the components on which the failed process depends, and the utility services (electrical or the supply of water, for example) on which these technical services depend, and the functional states of the physical assets involved.

Application of the methods extend beyond the integration of single utility networks to the modelling of interdependencies between multiple utilities and between these utilities and other infrastructure. The inference method of section 4.2 is limited, in part, by the partial coverage of urban features that influence design decisions and lack of standardised expression of spatial uncertainties (see section 5.2.4). Despite the extensive interdependencies of their services and initiatives such as the National Underground Asset Register (NUAR) (Geospatial Commission, 2019), utility providers do not currently have easy access to each other's datasets (see section 2.2.4 for examples). Data privacy, commercial sensitivity and security concerns – particularly for critical national infrastructure – impinge on exploitability of information-rich datasets (Geospatial Commission, 2020a). Once accessed, trust in the fidelity of the datasets becomes an issue. For underground utilities and the utility strike use case (see section 5.3.1), the accuracy of recorded location of an asset has significant implications on the risk of strike during excavation. Furthermore, the confidence with which inference methods (such as that demonstrated in Chapter 4) can be applied to decision-making depends on the availability and fidelity of digital representations. There is extensive scope for meaningful application of heuristics and inference (along with more deductive methods) to the large-scale modelling of real-time interactions between utility, transport and communications infrastructure if data standards are harmonised to the extent that datasets can be integrated to a sufficient depth – again, there is emphasis on minimum requirements for target use cases, which need to be defined through further research.

As discussed in 4.2.8, existing building stock has low BIM coverage but this is expected to change as the benefits of BIM are recognised more broadly and if BIM requirements on government-funded projects trickle across to private-sector projects. The case study described in Chapter 4 was limited to the Helix buildings but the increased use of BIM for new-builds and potential of retrospective BIM for existing stock (this will be more difficult for hidden MEP networks than visible architectural detail) present the possibility of scaling such studies to larger urban areas. With broader data coverage, the ability to study the impact on entire end-to-end networks of specific configuration choices for distribution networks would be realistic. However, section 4.3.5 also discussed the importance of a multi-factor approach to simulating truly representative real-world scenarios.

Real-time usage dynamics are needed for simulations and analyses to be truly representative of real-world scenarios. Just as digital modelling of static urban infrastructure is growing, it is expected that detailed, real-time monitoring of facilities is likely to become less exclusive to sophisticated research centre such as the USB (see section 3.2) and eventually become standard within both commercial and residential premises. In 2012, UK legislation started the smart meter rollout across the country with over 12 million homes being smart metered for electricity (and over 5 million for gas) by the end of 2018 and a goal of complete coverage by the end of 2020 (although smart meters are not obligatory) (Webborn *et al.*, 2019). There are already initiatives that make use of bulk collection and analysis of such data for research purposes, such as the Smart Energy Research Lab (SERL) of University College London (UCL) (SERL, 2021). Smart metering for water is also either in operation or planned for rollout over the next five years in many areas (Northumbrian Water Group, 2019; Thames Water, 2021). Although the granularity of residential smart metering – usually at or just inside the building envelope – may be sufficient for many urban planning use cases, there is also value in knowing the more granular breakdown of consumption inside buildings, such as the type of appliance that is consuming electricity, water or gas (or all of them). This is directly relevant to the use case detailed in section 2.2.3, for which a hospital was used as an example. If data describing the quantity, timing and *purpose* of water consumption (at that time) of a facility of known type is available, and if a supply interruption forces an intervention by the provider, the WDN could be dynamically reconfigured in such a way as to avoid an outage during a time-critical, high priority event. The water company needs to determine how to intervene to solve a problem while minimising the knock-on negative impact of this intervention.

For understanding fine-scale internal building consumption, an avenue of research concerns the probabilistic assignment of sensor data streams to physical asset based on non-intrusive load monitoring (NILM). Computers, lights, machinery, taps and other consumer appliances within buildings exhibit load signatures (LS) that can be detected and used to disaggregate consumption down to the appliance level. The benefits of NILM include enabling consumers to understand and potentially adapt their consumption behavior to reduce their costs (Bouhouras *et al.*, 2017). An ability to desegregate through signal decomposition could be used to address the issue of sparsely monitored facilities (section 5.3.3) and the attribution of monitoring data to physical assets and their digital element (section 5.2.5). Knowledge of the type of appliance within a physical space (from BIM element types, for example) and the type of appliance to which a sensor data stream relates (from NILM) could be used to allocate data streams to building appliances, potentially circumventing the need for explicit referencing of the monitored appliance within the data stream – this will depend on the detail of attribution of the BIM elements, complexity of usage through a single sensor and spatial granularity of sensing.

An alternative approach to filling gaps in coverage of user consumption is data emulation. A representative subset of real streaming data for a well understood facility type could be used to generate statistically valid but fictitious streams of consumption data as surrogates for real-time data feeds. And if BIM models for a facility also do not exist but its type is known, other spatial data such as building footprints could be used to weight emulated data by, for example, ground coverage as a proxy for consumption magnitude – an alternative heuristic use of building footprints to that used in section 4.2. Such statistical representation of usage patterns is also applicable to longer-term planning scenarios. BIM models could be used to select appropriate data stream emulators for the various buildings in a planned housing complex or industrial estate, and utility providers could use these data streams alongside the location of the (georeferenced) BIM models to locate this predicted future demand and simulate the redundancy offered (and cost incurred) by competing layout options for the distribution network that must be placed to service the buildings.

There is scope for applying partial 'direct' monitoring data, NILM and emulators to study the evolution of space and appliance usage under changing circumstances. For example, had all these methods been used to output a pre-Covid 19 simulation of usage in a building at a spatial granularity of rooms and a thematic granularity of activity, a post-Covid 19 model rebuild could provide insights into changes of usage patterns as a consequence of legislation introduced to deal with the pandemic. Such information could be used to reassess the suitability of temperature, lighting, air controls, ingress/egress options and access controls, and potentially to predict future behavioural outcomes of policies that are set in response to viral outbreaks.

If multiple infrastructure networks across urban and internal building scales could be integrated for the representation of interdependencies through an entire end-to-end system, the impact of infrastructure planning options on individual consumers could be assessed in greater detail through the simulation of cascading failures[37] down to the consumer end-points; this could include the fine-scale impact of utility network outages on transport and communication networks, and even the impact of these disruptions on behaviours in social networks (and vice versa). Further insights could be gleaned from integration of environmental and other sustainability factors: what are the local and regional environmental consequences of these internal building design and engineering options, both short- and long-term? Will this project choice incur more destruction of natural habitat and the generation of more air pollution? Questions such as these are well beyond the scope of utility or even broader infrastructure network integration but concern meaningful consequences of relevant decisions; consequences that are difficult to (or should not be) monetised or quantified in terms of operational efficiency or reliability.

[37] Cascading failures of coarser scale critical national infrastructure networks are studied in detail in Craig Robson's PhD thesis (Robson, 2016)

Chapter 6 Conclusions

6.1 Introduction

This thesis has demonstrated that information abstraction, graph representations and spatial inference algorithms can be used to construct integrated representations of the topology of utility networks across the building envelope. This thesis has also demonstrated that improvements to data standardisation and best practice in the implementation of these standards is important for these integration processes to be more automated. There must be mechanisms that enable data producers to express inter-dataset connectivity, commonly and uniquely identify real-world objects, and provide absolute location of utility assets (with meaningful expression of uncertainties) that is agreed and standardised across all built environment domains. The remainder of this chapter highlights the key points of discussion on the research in this thesis that supports this statement (section 6.2), the novelty and contributions of the research (section 6.3), the wider implication of this research and opportunities for future work that would build on the these contributions (section 6.4).

6.2 Summary of key discussion points

The findability, accessibility and coverage of datasets proved challenging to the research. Whether data exists, where and how they can be sourced, and the presence of commercial sensitivities and security constraints are significant factors in the achievability of data integration objectives. The discovery of and access to digital representations of the assets that connect buildings to distribution networks and the lack of georeferencing in the sourced data are problematic and important issues. Testing the scalability of the methods is inhibited by a lack of detailed BIM modelling (especially detailed MEP data) for most of the building stock. It is suggested that an expansion to the normal, practical spatial remit of high-detail BIM modelling to the area between buildings and distribution networks (in place of non-georeferenced CAD data) would increase the coverage of georeferenced, machine-readable representations of utility networks.

Consistent and accurate georeferencing of BIM data enables the inference of intra-dataset network topologies in support of some key use cases. However, the inconsistent and sometimes inaccurate use of georeferencing in BIM models – and the inability to automatically corroborate the coordinates – inhibits the geospatial exploitability and hence scope of the practical applicability of the datasets. This is partly caused by the lack of full or correct data attribution by data producers and partly by inadequate data standardisation. IFC needs to be modified to better support georeferencing and the enablement of internal-dataset corroboration of georeferences, providing users with a means of asserting confidence in the location and orientation of projects without the need for burdensome manual inspection. Across both the geospatial and BIM domains, there is a general problem with inconsistency of data precision and accuracy, with the use of fine precision (often to the nearest millimetre) implying falsely high accuracy; there is no cross-domain standardisation for the expressing of numerical uncertainty, resulting in unnecessary 'unknown unknowns'; inconsistency in chosen spatial units and the identification of these units incurred manual intervention; and a general low accuracy of location data for utility assets limits the fineness of scale at which spatial inference algorithms can be applied with meaningful confidence.

Lack of explicit connectivity between the datasets that provide digital representations of utility networks is a fundamental problem and the basis for much of the research of this thesis. Generally, utility network datasets do not contain attributes that enable a linking to other datasets that represent subnetworks to which the subject one connects; many other observations, conclusions and suggestions in this thesis are intended to circumvent this deficiency, which is very challenging to resolve in practice. The challenge is due partly to a lack of complete digital coverage of utility assets in the zone between the building envelope and distribution networks but it is also due to the frequent absence of absolute location data in these datasets. Regardless of dataset coverage and location data, for confident and reliable assertion of intra-dataset connectivity, there is a need for data attributes that allow the referencing of real-world assets that may be present in other datasets but the effective use of such attributes is limited by not knowing what other datasets or assets exist and the absence of a standardised way of identifying them. Standards should enable the attribution of asset flow position within network hierarchies and whether the asset is expected to connect to another sub-network in the real world; along with more accurate location data, this would increase confidence in the outputs of inference methods.

The challenge of uniquely and commonly identifying real-world objects in the built environment is an important issue for automated integration of utility network topology and digital twin challenges more broadly; the key point is ensuring that there is one identifier that is (or can be) used commonly by all digital representations across all built environment domains, and it should be possible to identify these objects even before their construction is complete and before they are represented digitally. In practice, achieving this is difficult and it is recommended that future research considers possible approaches. It is proposed that a feasible solution for generating such common, universally unique identifiers (CUUIDs) could involve the concatenation of object type and absolute coordinate but would likely require the development of a standardised dictionary of common concepts and the use of an agreed coordinate references system. Proprietary object attribution is identified specifically as a data-related challenge that inhibits integration for some target use cases.

The ability to assert correspondence between digital representation of assets and sensor (or monitoring) data is emphasised as a capability of high value to utility resource flow simulations and digital twins more broadly. BIM mechanical, electrical and plumbing (MEP) elements tend not to reference (building management system) BMS data streams or their sensors, nor vice versa. Furthermore, internal building monitoring data is usually not accompanied by spatial data, such that it is difficult to infer associations where they are not explicit. Integration of IoT and BIM is still in its early stages and is a research topic that merits exploration.

This technical work of this thesis finishes with a study on some performance characteristics of water distribution network (WDN) configurations using graph theory. Consequently, several domain-specific research insights are specific to potable water supply networks. It is shown that BIM MEP models and WDNs can be used in support of graph theory-based assessments of network robustness and redundancy for different BIM-WDN layout options but practical applicability is limited by the questionable usefulness of knowing optimal spatial-topological configurations of entire urban networks (rather than the benefit to the local areas of interest to developers) and by the 'dead-end' nature of consumer premises – sections of a WDN are not connected by buildings, such that their placement within the WDN has little impact on whole-network robustness or redundancy. It is suggested that buildings could serve as supply 'bridges' between areas of a WDN that offer independent supply as a means of increasing redundancy; in this context, optimisation of BIM-WDN configuration would have higher impact on supply reliability. However, there would be liability implications of such service bridges: who would be responsible for continuation of supply through the connections and held accountable in the event of flow disruption? A further consideration is that the computational complexity of calculating spectral graph theory metrics for competing BIM-WDN configuration options over large urban areas is potential prohibitively high (due to the high number of iterations) but this should be further tested. It is concluded that physics-based, hydraulic modelling and real-time usage data should be used to identify the flow-connected portions of a WDN that are to be included in graph theory calculations, but that a lack of integration and complete overlap of geospatial and hydraulic models prevents the realistic weighting of the matrices used in spectral graph theory metrics. There is scope for using graph theory metric for routing optimisation but this should be supported by physics-based modelling.

The Integrated Digital Built Environment (IDBE) working group's action points at the end of their discussion document (Gilbert *et al.*, 2020) concern best practice and data standardisation within the building environment; these points overlap and are consistent with many of the findings from this thesis. There are several future capabilities that would be enabled by further development of the methods of this thesis and more extensive deployment of the technologies explored. Many of these capabilities depend on addressing some of the above data standardisation and best practice considerations. Although this thesis alone cannot be used as the only source for what is and should remain a consensus-based standardisation process, it provides some initial insights with relevance to utility networks, which could be used to inform or seed discussions around the development of existing standards such as IFC and the CityGML UtilityNetwork ADE.

Inference of the presence and location of underground utility assets for avoidance of accidental strike is a strong use case for the spatial inference algorithms developed in this thesis. Additional research needs to consider the integration of other underground features and structures in the heuristics. There is strong focus in UK and internationally on the value of accurate and complete digital representation of underground utility assets and the sharing of datasets between utility providers. The Geospatial Commission's National Underground Asset Register (NUAR) project and the development of the Model for Underground Data Definition and Integration (MUDDI) data standard under the open Geospatial Consortium are both testament to this importance.

Graph database and message broker technologies are assessed as having the potential to address some of the challenges facing the digital twin domain by enabling the dynamic representation of the structure of and flows through infrastructure networks, especially when the subject datasets are underpinned by data standards exhibiting different relationship structures and conceptualisation. It is concluded that graph representations and graph databases can be used effectively for the integration of disparate and disjoint urban datasets, and for the representation in real-time of electricity consumption across multiple spatial scales. Publish/subscribe message brokers are an effective means of communicating dynamic network structures and flow states to multiple consumer systems.

6.3 Novelty and contributions

This thesis defines and then uses as guidance a set of domain-specific use cases in support of evaluating of what is fundamentally required of integration methods. These use cases were allowed to evolve through focussing on requirements from industry and close collaboration with standards bodies. Requirements were gathered from discussions with a utility company, an estates management department and multiple construction contractors; the discussions revolved around in-depth case studies that provided the content and tangibility needed for grounded redirection of research that was carried out from a position of impartiality. This approach strengthened the research impact and is recommended for future academic research that is intended to influence the development of data standards.

There appears to be no consistent understanding of what is meant by ‘integration’ and ‘interoperability’ within the built environment domain, although there is a general tendency to imply something that relates to harmony of data schemas, simultaneous visualisation or both. The thesis defines explicitly one interpretation of 'integration' in the context of utility networks: the process of asserting connectivity between multiple disjoint digital representations of real-world sub-networks that are physically and functionally connected in the real-world, allowing them to be analysed as a single network in a digital environment. It is this interpretation that led the research to reinforce a growing consensus that partial information retention that is sufficient for targeted use cases is more practical than complete integration that depends on lossless data conversions.

Spatial inference and heuristics were used in this thesis to fill gaps in partial digital representations and assert connectivity between subnetworks represented by the different datasets. This probabilistic approach to integration of network connectivity is novel and there is potential to exploit a wider variety of datasets than those generally used for digital representation of the built environment; an example is the use of synthetic aperture radar interferometry for the measurement of subtle ground movement that, without intervention, can damage underground assets and cause utility supply failure.

This thesis proposes the concept of a universally unique identifier that is common to all digital representations and can be attributed before a physical object is represented digitally or even constructed in the real world; it is suggested that the basis of such an identifier could be, for example, the concatenation of an object type value (from an agreed set of object types) and a real-world coordinate expressed in an agreed coordinate reference system.

This thesis evidences the potential benefits of a pragmatic and incremental approach to built-environment data standardisation; it shows that specific use cases can be well supported through adjustments to existing data standards. It has also demonstrated an incentive for the public sector to stimulate – and potentially fund – the enrichment of urban datasets for use cases that are of longer-term public benefit. With businesses under pressure to return profits within the shorter-term timescales of individual projects, public oversight might be needed to protect the longer-term, national-level objectives concerning digital twins and smart cities.

6.4 Wider implications and future work

The MUDDI standard is relatively new, still under initial development and could align with future use cases for projects such as the Geospatial Commission's NUAR project. Ordnance Survey is a sponsor of this research and has a stake in both MUDDI and NUAR. Although NUAR is currently focussed on facilitating the sharing of asset location data between utility companies, if its remit does expand to encompass multi-utility dependency modelling, it would stand to benefit from an ability to confidently derive connections between utility types – these could include utilities beyond the consideration of this thesis, such as fibre-optic communication. The alignment of MUDDI with some recommendations in this theses could support methods of spatial inference that can be used to represent such multi-utility coupling; in particular, the standardisation of representations of spatial uncertainty would support this.

It is suggested that the AEC and utilities industries consider the use of building-to-building utility 'bridge' connections, adding flexibility to the reconfigurability of network topologies and thus increasing supply reliability. Allowing facilities that are supplied from different sections of a network to support each other in the event of supply failure via a normal routing would likely be impactful for critical facilities such as hospitals.

An implication for the public sector is that it might need to take responsibility of incentivising financially some longer-term investment in data quality or richness. Many of the issues concerning the data standardisation and best practices issues described in Chapter 5 might be best addressed initially through legislation and public financing. Given the likely longer-term benefits and associated financial risk with (for example) adherence to public data standards, full richness of dataset attribution (beyond immediate project requirements) and the following of renewed best practice. Given the privacy, commercial sensitivity and security risks associated with implementing some of the suggestions for critical national infrastructure, government legislation and funding might be needed to provide sufficient incentives and protection for the suggestions to be practical, profitable and sustainable.

Future research might involve experimenting with a combined use of the technologies used in Chapter 3 with the heuristic inference algorithms developed for Chapter 4, and then applying this to more complex problems. For example, an integration of graph databases, message brokers, heuristics and inference (along with more deductive methods) could be applied to the large-scale modelling of real-time interactions between utility, transport and communications infrastructure. Such systems could comprise important components of cities' digital twins. Such a modelling system would enable studies on the impact on entire multi-utility, end-to-end networks of specific configuration choices for distribution networks. However, the feasibility of this likely depends on addressing some of data challenges described in the discussion. In particular, the sparsity of sensor data coverage, discontinuities within BIM MEP network topology and issues concerning useful data coverage between the building envelope and distribution networks should be addressed as a next step in follow-on research.

In the longer-term, a combined use of inference from non-intrusive load monitoring and plausible emulation could address the sparsity of direct monitoring data in the built environment. The technologies and methods have applicability to the modelling of multi-utility and multi-infrastructure interdependencies and detailed, fine scale impact of various scenarios and decision-making in the built environment, including cascading failure simulations and studies of feedback loops from social behaviours and environmental impacts.

References

Abdelbaki, C. *et al.* (2017) 'Management of a water distribution network by coupling GIS and hydraulic modeling: a case study of Chetouane in Algeria', *Applied Water Science*, 7(3), pp. 1561–1567. doi: 10.1007/s13201-016-0416-1.

Affinity Water (2018) *Affinity Water Limited Developer Services Guidance on Self-Lay Design , Permissible Materials and Construction Arrangements*. Available at: https://www.affinitywater.co.uk/docs/developer/Guidance-on-Self-Lay.pdf.

Affinity Water (2019) *Developing and working near our pipes and apparatus*. Available at: https://www.affinitywater.co.uk/docs/developer/Building-Near-Pipes-Apparatus-Guide-17-04-2019-final.pdf.

Agathokleous, A., Christodoulou, C. and Christodoulou, S. E. (2017) 'Topological Robustness and Vulnerability Assessment of Water Distribution Networks', *Water Resources Management*, 31, pp. 4007–4021. doi: 10.1007/s11269-017-1721-7.

Ahmed, N. *et al.* (2018) 'Residential Consumer-Centric Demand Side Management', 9(5), pp. 4513–4524.

Ahopelto, S. and Vahala, R. (2020) 'Cost-Benefit Analysis of Leakage Reduction Methods in Water Supply Networks', *Water (Switzerland)*, 12(1). doi: 10.3390/w12010195.

Akinci, B. *et al.* (2008) 'CAD and GIS interoperability through semantic web services', *Electronic Journal of Information Technology in Construction*, 13.

Akinci, Burcu *et al.* (2008) 'CAD and GIS interoperability through semantic web services', *Electronic Journal of Information Technology in Construction*, 13(December), pp. 39–55. doi: 10.1201/9781420068061-c9.

de Almeida, J. P., Morley, J. G. and Dowman, I. J. (2013) 'A graph-based algorithm to define urban topology from unstructured geospatial data', *International Journal of Geographical Information Science*, 27(8), pp. 1514–1529. doi: 10.1080/13658816.2012.756881.

Amaral, L. A. N., Scala, A. and Barthe, M. (2000) 'Classes of small-world networks', 97(21). doi: 10.1073/pnas.200327197.

Amarasinghe, P. (2014) *RESILIENCE OF WATER SUPPLY SYSTEMS IN MEETING THE CHALLENGES POSED BY CLIMATE CHANGE AND POPULATION GROWTH*. Available at: https://core.ac.uk/download/pdf/33496023.pdf (Accessed: 24 August 2020).

Amber, S. (2011) *Examining the Agile Manifesto*. Available at: http://www.ambysoft.com/essays/agileManifesto.html (Accessed: 23 December 2020).

Amirebrahimi, S. *et al.* (2015a) 'A data model for integrating GIS and BIM for assessment and 3D visualisation of flood damage to building', in *CEUR Workshop Proceedings*, pp. 78–89.

Amirebrahimi, S. *et al.* (2015b) 'A framework for a microscale flood damage assessment and visualization for a building using BIM–GIS integration', *International Journal of Digital Earth*, 8947(April 2015), pp. 1–24. doi: 10.1080/17538947.2015.1034201.

Amirebrahimi, S. *et al.* (2016) 'A BIM-GIS integration method in support of the assessment and 3D visualisation of flood damage to a building', *Journal of Spatial Science*, 61(2), pp. 317–350. doi: 10.1080/14498596.2016.1189365.

Andrews, C. (2020) *Common Patterns for BIM and GIS Integration, ArcGIS Blog*. Available at: https://www.esri.com/arcgis-blog/products/arcgis-pro/transportation/common-patterns-for-bim-and-gis-integration/ (Accessed: 17 November 2020).

Apache Foundation (2020) *Apache Kafka*. Available at: https://kafka.apache.org/ (Accessed: 20 March 2018).

Arayici, Y. (2007) 'Towards building information modelling for existing structures', *Structural Survey*, 26(3), pp. 210–222. doi: 10.1108/02630800810887108.

Aşchilean, I. *et al.* (2018) 'The unfavourable impact of street traffic on water distribution pipelines', *Water (Switzerland)*, 10(8), pp. 1–18. doi: 10.3390/w10081086.

ASHRAE (2020) *BACnet.org*. Available at: http://www.bacnet.org/ (Accessed: 18 February 2021).

ASPRS (2014) 'ASPRS Positional Accuracy Standards for Digital Geospatial Data', *Photogrammetric Engineering & Remote Sensing*, 81(3), pp. 1–26. doi: 10.14358/pers.81.3.a1-a26.

Atkinson, S. (2012) 'Design guidance for water mains and services on new development sites', (3). Available at: http://www.unitedutilities.com/documents/DesignManualIssueMay2012.pdf.

Aussel, D. *et al.* (2020) 'A trilevel model for best response in energy demand-side management', *European Journal of Operational Research*, 281(2), pp. 299–315. doi: 10.1016/j.ejor.2019.03.005.

Autodesk (2020) *Navisworks, Autodesk Knowledge Network*. Available at: https://knowledge.autodesk.com/support/navisworks-products/learn-explore/caas/CloudHelp/cloudhelp/2020/ENU/Navisworks/files/GUID-983300A3-31FC-420C-83C9-4AEEDBE5E083-htm.html (Accessed: 18 November 2020).

Awad, H., Yassin, A. and Ayad, A. (2017) 'Geographic Information Systems in Water Distribution Networks', (June), p. 14.

Azavea (2021) *Open Civic Data Standards*. Available at: https://azavea.gitbooks.io/open-data-standards/content/ (Accessed: 1 March 2021).

Azevedo, B. B. and Saurin, T. A. (2018) 'Losses in Water Distribution Systems: A Complexity Theory Perspective', *Water Resources Management*, 32(9), pp. 2919–2936. doi: 10.1007/s11269-018-1976-7.

Baccino, F. *et al.* (2013) 'Domestic heat load aggregation strategies for wind following in electric distribution systems', *IEEE Power and Energy Society General Meeting*. doi: 10.1109/PESMG.2013.6672610.

Baldick, R. *et al.* (2009) 'Vulnerability assessment for cascading failures in electric power systems', *2009 IEEE/PES Power Systems Conference and Exposition, PSCE 2009*, pp. 1–9. doi: 10.1109/PSCE.2009.4839939.

Bari, P. (2015) 'Smart Cities : Definitions , Dimensions , Performance , and Initiatives', (April). doi: 10.1080/10630732.2014.942092.

Barnes, P. and Davies, N. (2014) *BIM in Principle and in Practice*. ICE Publishing. doi: doi:10.1680/bimpp.60920.

Batty, M. and Hudson-Smith, A. (2006) 'Digital cornucopias: Changing conceptions of the virtual city', *Environment & Planning B*, 33(6), pp. 799–802. doi: 10.1068/b3306ed.

BBC (2013) *Welshpool hospital water outage*. Available at: https://www.bbc.co.uk/news/uk-wales-mid-wales-24080554 (Accessed: 13 November 2021).

Beck, K. *et al.* (2010) *Manifesto for Agile Software Development*. Available at: http://agilemanifesto.org/ (Accessed: 23 December 2020).

Becker, T., Nagel, C. and Kolbe, T. H. (2010) 'UtilityNetworkADE core model'.

Becker, T., Nagel, C. and Kolbe, T. H. (2011) 'Integrated 3D Modeling of Multi-utility Networks and Their Interdependencies for Critical Infrastructure Analysis', *Advances in 3D Geo-Information Sciences Lecture Notes in Geoinformation and Cartography*, pp. 73–93. doi: 10.1007/978-3-642-12670-3_1.

Becker, T., Nagel, C. and Kolbe, T. H. (2013) 'Semantic 3D modeling of multi-utility networks in cities for analysis and 3D visualization', *Lecture Notes in Geoinformation and Cartography*, pp. 41–62. doi: 10.1007/978-3-642-29793-9_3.

Becker, T., Nagel, C. and Kolbe, T. H. (2016) *UtilityNetworkADE*. Available at: http://www.citygmlwiki.org/index.php/CityGML_UtilityNetworkADE (Accessed: 15 September 2017).

Beetz, J., van Leeuwen, J. and de Vries, B. (2009) 'IfcOWL: A case of transforming EXPRESS schemas into ontologies', *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 23(01), p. 89. doi: 10.1017/S0890060409000122.

Benner, J *et al.* (2005) 'Flexible Generation of Semantic 3D Building Models', *Workshop on Next Generation 3D City Models*, 49, pp. 17–22.

Bernhardsen, T. (2002) *Geographic Information Systems: An Introduction*. Third. John Wiley & Sons.

Biljecki, F., Kumar, K. and Nagel, C. (2018) 'CityGML Application Domain Extension (ADE): overview of developments', 3(13). doi: 10.1186/s40965-018-0055-6.

Bishr, Y. and Kuhn, W. (2000) 'Ontology-Based Modelling of Geospatial Information Role of Ontologies in Information Sharing Ontology and Conceptual Modelling', pp. 25–28.

Boccaletti, S. *et al.* (2006) 'Complex networks: Structure and dynamics', *Physics Reports*, 424(4–5), pp. 175–308. doi: 10.1016/j.physrep.2005.10.009.

Bolton, A. *et al.* (2018) *The Gemini Principles, Centre for Digital Built Britain*. doi: 10.17863/CAM.32260.

Borrmann, A. *et al.* (2015) 'Multi-scale geometric-semantic modeling of shield tunnels for GIS and BIM applications', *Computer-Aided Civil and Infrastructure Engineering*, 30, pp. 263–281. doi: 10.1111/mice.12090.

Bostock, M. (2018) *d3 JavaScript library*. Available at: https://d3js.org/.

Bostock, M. (2019) *d3-Sankey plugin*. Available at: https://github.com/d3/d3-sankey (Accessed: 9 April 2018).

Bouhouras, A. S. *et al.* (2017) 'Load signature formulation for non-intrusive load monitoring based on current measurements', *Energies*, 10(4). doi: 10.3390/en10040538.

Bradley, A. *et al.* (2016) 'BIM for infrastructure: An overall review and constructor perspective', *Automation in Construction*. doi: 10.1016/j.autcon.2016.08.019.

Briscoe, D. (2016) *Beyond BIM: Architecture Information Modeling*. Routledge.

British Standard Institution (2014) *BS 1192-4 : 2014 Collaborative production of information Part 4 : Fulfilling employer' s information exchange requirements using COBie - Code of practice, British Standards Institution*. Available at: http://shop.bsigroup.com/forms/BS-1192-4/.

British Standards Institution (2014a) *PAS 180:2014 Smart cities – Vocabulary, British Standard Institution*.

British Standards Institution (2014b) *PAS 181: Smart City Framework – Guide to establishing strategies for smart cities and communities, British Standard Institution*.

British Standards Institution (2014c) *PAS 182:2014 Smart city concept model*.

British Standards Institution (2017) *PAS 183: 2017 Smart cities – Guide to establishing a decision-making framework for sharing data and information services*.

BSI Group (2021) *ISO 19650*. Available at: https://www.bsigroup.com/en-GB/iso-19650-BIM/ (Accessed: 3 March 2021).

Bu, S., Yu, F. R. and Liu, P. X. (2011) 'Dynamic pricing for demand-side management in the smart grid', *2011 IEEE Online Conference on Green Communications, GreenCom'11*, pp. 47–51. doi: 10.1109/GreenCom.2011.6082506.

buildingSMART (2020) *buildingSmart Data Dictionary*. Available at: https://www.buildingsmart.org/users/services/buildingsmart-data-dictionary/ (Accessed: 2 November 2020).

buildingSMART (2021) *Industry Foundation Classes*. Available at: https://www.buildingsmart.org/standards/bsi-standards/industry-foundation-classes/ (Accessed: 9 March 2021).

buildingSMART alliance (2015) 'NBIMS-US-3: 1 - Scope', *National BIM Standard - United States - Version 3*, pp. 1–5.

buildingSMART International (2020a) *IFC, bSI*. Available at: https://www.buildingsmart.org/standards/bsi-standards/industry-foundation-classes/ (Accessed: 17 June 2020).

buildingSMART International (2020b) *openBIM*. Available at: https://www.buildingsmart.org/about/openbim/openbim-definition/ (Accessed: 10 November 2020).

BuildingSMART and Leibich, T. (2009) *IFC 2x Edition 3: Model Implementation Guide*, *System*.

Butler, D. *et al.* (2017) 'Reliable, resilient and sustainable water management: the Safe & SuRe approach', *Global Challenges*, 1(1), pp. 63–77. doi: 10.1002/gch2.1010.

Cabinet Office (2018) 'Geospatial Economic Opportunity Initial Analysis : Geospatial Opportunity', (August), pp. 1–18.

Cabinet Office, H. (2011) 'Government Construction Strategy', *Construction*, 96(May), p. 43. doi: Vol 19.

Cameron, D. A. (2001) ‘The extent of soil desiccation near trees in a semi-arid environment’, *Geotechnical and Geological Engineering*, 19(3–4), pp. 357–370. doi: 10.1023/A:1013168708654.

Centre for Digital Built Britain (2020) ‘The approach to delivering a National Digital Twin for the United Kingdom’. Available at: https://www.cdbb.cam.ac.uk/files/approach_summaryreport_final.pdf.

Charalambous, B. (2008) ‘Use of district metered areas coupled with pressure optimisation to reduce leakage’, *Water Science and Technology: Water Supply*, 8(1), pp. 57–62. doi: 10.2166/ws.2008.030.

Chen, H.-M., Chang, K.-C. and Lin, T.-H. (2016) ‘A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs’. doi: 10.1016/j.autcon.2016.03.002.

Cheng, J. C. P., Deng, Y. and Anumba, C. (2015) ‘Mapping BIM schema and 3D GIS schema semi-automatically utilizing linguistic and text mining techniques’, *Journal of Information Technology in Construction*, 20.

Chi, H. L., Kang, S. C. and Wang, X. (2013) ‘Research trends and opportunities of augmented reality applications in architecture, engineering, and construction’, *Automation in Construction*, 33, pp. 116–122. doi: 10.1016/j.autcon.2012.12.017.

CityJSON contributors (2019) *CityJSON*. Available at: https://www.cityjson.org/ (Accessed: 26 February 2021).

Cuenca, J., Larrinaga, F. and Curry, E. (2017) ‘A unified semantic ontology for energy management applications’, *CEUR Workshop Proceedings*, 1936, pp. 86–97.

Dataincubators (2012) *Smushing*. Available at: https://patterns.dataincubator.org/book/smushing.html (Accessed: 7 November 2020).

Daum, S. and Borrmann, A. (2014) ‘Processing of topological BIM queries using boundary representation based methods’, *Advanced Engineering Informatics*, 28(4), pp. 272–286. doi: 10.1016/j.aei.2014.06.001.

Dave, B. *et al.* (2018) ‘A framework for integrating BIM and IoT through open standards’, *Automation in Construction*, 95(August), pp. 35–45. doi: 10.1016/j.autcon.2018.07.022.

Deep Dig Output (2017) *Underground Assets Challenge*. Available at: https://geovation.uk/wp-content/uploads/2017/01/A00031-Deep-Dig-Booklet_UPDATE.pdf (Accessed: 21 November 2020).

Deleuran, A. H. and Derix, C. (2013) 'Topological Infrastructure Analysis of the Built Environment', *EAEA-11 conference 2013 - Conceptual Representation: exploring the layout of the built environment*, (Track 3), pp. 419–426.

Deng, Y., Cheng, J. C. P. and Anumba, C. (2016) 'Mapping between BIM and 3D GIS in different levels of detail using schema mediation and instance comparison', *Automation in Construction*. doi: 10.1016/j.autcon.2016.03.006.

Desai, K. R. and Dutta, G. (2013) 'A dynamic pricing approach to electricity prices in the Indian context', *International Journal of Revenue Management*. doi: 10.1504/IJRM.2013.059625.

Diakite, A. A. and Zlatanova, S. (2020) 'Automatic geo-referencing of BIM in GIS environments using building footprints', *Computers, Environment and Urban Systems*, 80. doi: 10.1016/j.compenvurbsys.2019.101453.

Ding, L. Y. *et al.* (2016) 'Construction risk knowledge management in BIM using ontology and semantic web technology'. doi: 10.1016/j.ssci.2016.04.008.

DNSstuff Staff (2019) *What is a network topology?, DNSstuff*. Available at: https://www.dnsstuff.com/what-is-network-topology (Accessed: 12 November 2020).

Döllner, J. and Hagedorn, B. (2008) 'Integrating urban GIS, CAD, and BIM data by service-based virtual 3D city models', in *Proceedings of the Urban and Regional Data Management - UDMS Annual 2007*, pp. 157–170. doi: 10.1017/CBO9781107415324.004.

Donkers, S. *et al.* (2016) 'Automatic conversion of IFC datasets to geometrically and semantically correct CityGML LOD3 buildings', *Transactions in GIS*, 20(4), pp. 547–569. doi: 10.1111/tgis.12162.

Dore, C. and Murphy, M. (2012) 'Integration of Historic Building Information Modeling (HBIM) and 3D GIS for recording and managing cultural heritage sites', in *Proceedings of* *the 2012 18th International Conference on Virtual Systems and Multimedia, VSMM 2012: Virtual Systems in the Information Society*, pp. 369–376. doi: 10.1109/VSMM.2012.6365947.

Dutta, G. and Mitra, K. (2017) ‘A literature review on dynamic pricing of electricity’, *Journal of the Operational Research Society*, 68(10), pp. 1131–1145. doi: 10.1057/s41274-016-0149-4.

East, E. W. (2007) ‘Construction Operations Building Information Exchange (COBIE): Requirements Definition and Pilot Implementation Standard’, *Construction Engineering Research Laboratory ERDC/CERL TR-07-30*, (June), pp. 1–195.

Eastman, C. *et al.* (2011) *A guide to building information modeling*. Second. John Wiley & Sons.

El-Mekawy, M. (2010) *Integrating BIM and GIS for 3D City Modelling: The Case of IFC and CityGML*, *Advances in 3D Geo-Information Sciences*. doi: urn:nbn:se:kth:diva-28899.

El-Mekawy, M., Östman, Anders and Hijazi, I. (2012) ‘A Unified Building Model for 3D Urban GIS’, *ISPRS International Journal of Geo-Information*, 1(2), pp. 120–145. doi: 10.3390/ijgi1020120.

El-Mekawy, M., Östman, A and Hijazi, I. (2012) ‘An Evaluation of IFC-CityGML Unidirectional Conversion’, *International Journal of Advanced Computer Science and Applications*, 3(5), pp. 159–171. doi: 10.14569/IJACSA.2012.030525.

El-Mekawy, M., Ostman, A. and Shahzad, K. (2011) ‘Towards Interoperating CityGML and IFC Building Models: A Unified Model Based Approach’, *Advances in 3D Geo-Information Sciences Lecture Notes in Geoinformation and Cartography*, pp. 73–93. doi: 10.1007/978-3-642-12670-3.

Emmer, M. (2013) ‘Topology from Art and Math to contemporary cities’, *AusArt*, 1(1), pp. 255–264.

EPA (2019) *EPA Data Standards*. Available at: https://www.epa.gov/data-standards/learn-about-data-standards (Accessed: 11 June 2019).

Erdener, B. C. *et al.* (2014) ‘An integrated simulation model for analysing electricity and gas systems’, *International Journal of Electrical Power and Energy Systems*, 61, pp. 410–420.

doi: 10.1016/j.ijepes.2014.03.052.

Escobar, P. *et al.* (2020) 'An ontology-based framework for publishing and exploiting linked open data: A use case on water resources management', *Applied Sciences (Switzerland)*, 10(3). doi: 10.3390/app10030779.

Esekhaigbe, E., Kazan, E. and Usmen, M. (2020) 'Integration of Digital Technologies into Underground Utility Asset Management', *Open Journal of Civil Engineering*, 10(04), pp. 403–428. doi: 10.4236/ojce.2020.104030.

Esri (1999) 'Understanding ArcSDE ® : GIS by ESRI ®'. Available at: www.esri.com.

Estrada, E. (2006) 'Network robustness to targeted attacks. the interplay of expansibility and degree distribution', *European Physical Journal B*, 52(4), pp. 563–574. doi: 10.1140/epjb/e2006-00330-7.

European Commission (2012) 'D2.8.III.6 INSPIRE Data Specification on Utility and governmental services - Draft Technical Guidelines', (March), p. 233.

European Commission (2013) 'INSPIRE Metadata Implementing Rules : Technical Guidelines based on EN ISO 19115 and EN ISO 19119 - V.1.3.', p. 99. Available at: http://inspire.ec.europa.eu/.

Everflow Water (2019) *What is a SPID?* Available at: https://www.everflowwater.com/single-post/2019/11/21/what-is-a-spid (Accessed: 26 February 2021).

Falkowski, K. and Ebert, J. (2009) 'Graph-based urban object model processing', *and Traffic (CMRT'09): Object Extraction for 3D*, XXXVIII, pp. 3–4. Available at: http://www.pf.bv.tum.de/isprs/cmrt09/pub/CMRT09_Falkowski_Ebert.pdf.

Farooq, A. *et al.* (2018) 'Transportation planning through GIS and multicriteria analysis: Case study of Beijing and XiongAn', *Journal of Advanced Transportation*, 2018. doi: 10.1155/2018/2696037.

Feng, M. and Porter, M. A. (2020) 'Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence', *Physical Review Research*, 2(3). doi: 10.1103/physrevresearch.2.033426.

Ferrari, G., Savic, D. and Becciu, G. (2014) ‘Graph-theoretic approach and sound engineering principles for design of district metered areas’, *Journal of Water Resources Planning and Management*, 140(12), pp. 1–13. doi: 10.1061/(ASCE)WR.1943-5452.0000424.

Ferreira, N., Werner, H. and Park, M. (2015) ‘Urbane : A 3D Framework to Support Data Driven Decision Making in Urban Development’, pp. 97–104.

Fiedler, M. (1973) ‘Algebraic connectivity of graphs’, *Czechoslovak Mathematical Journal*, 23(2398), pp. 298–305. Available at: http://dml.cz/dmlcz/101168.

Fikes, R. and Farquhar, A. (1999) ‘Distributed Repositories of Highly Expressive Reusable Ontologies’, *IEEE Intelligent Systems and Their Applications*, 14(2), pp. 73–79. doi: 10.1109/5254.757634.

Finn, P. and Fitzpatrick, C. (2014) ‘Demand side management of industrial electricity consumption: Promoting the use of renewable energy through real-time pricing’, *Applied Energy*, 113(2014), pp. 11–21. doi: 10.1016/j.apenergy.2013.07.003.

Floros, G. S., Ellul, C. and Dimopoulou, E. (2018) ‘Investigating interoperability capabilities between IFC and CityGML LoD 4 - Retaining semantic information’, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(4/W10), pp. 33–40. doi: 10.5194/isprs-archives-XLII-4-W10-33-2018.

Fonseca, F., Davis, C. and Camara, G. (2003) ‘Bridging Ontologies and Conceptual Schemas in Geographic Applications Development’, *Geoinformatica*, 7(4), pp. 355–378.

Fonseca, F. and Martin, J. (2007) ‘Learning the differences between ontologies and conceptual schemas through ontology-driven information systems’, *Journal of the Association for Information Systems*, 8(2), pp. 129–142. doi: 10.17705/1jais.00114.

Fosu, R. *et al.* (2015) ‘Integration of Building Information Modeling (BIM) and Geographic Information Systems (GIS) – a literature review and future needs’, *Proc. of the 32nd CIB W78 Conference 2015, 27th-29th October 2015, Eindhoven, The Netherlands*, pp. 196–204.

Free CAD Web (2020) *Sept Solid*. Available at: https://wiki.freecadweb.org/Part_Sweep (Accessed: 24 September 2020).

GBFS contributors (2021a) ‘GBFS’. Available at: https://azavea.gitbooks.io/open-data-standards/content/standards/domain_specific_standards/general_bikeshare_feed.html.

GBFS contributors (2021b) *GBFS GitHub*. Available at: https://github.com/NABSA/gbfs (Accessed: 1 March 2021).

Geospatial Commission (2019) *Getting under the surface of our National Underground Assets Register (NUAR) team*. Available at: https://geospatialcommission.blog.gov.uk/2019/12/18/getting-under-the-surface-of-our-national-underground-assets-register-nuar-team/ (Accessed: 24 May 2020).

Geospatial Commission (2020a) *National Underground Asset Register*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/897060/Project_Orange_V3_01.07.pdf (Accessed: 17 August 2020).

Geospatial Commission (2020b) *Unlocking the Power of Location: The UK's geospatial strategy 2020 to 2025*. Available at: https://www.gov.uk/government/publications/unlocking-the-power-of-locationthe-uks-geospatial-strategy/unlocking-the-power-of-location-the-uks-geospatial-strategy-2020-to-2025 (Accessed: 15 September 2020).

Germanopoulos, G. and Jowitt, P. W. (1989) 'Leakage reduction by excess pressure minimization in a water supply network', *Proceedings - Institution of Civil Engineers. Part 2. Research and theory*, 87(August), pp. 195–214. doi: 10.1680/iicep.1989.2003.

Gilbert, T. *et al.* (2020) *Built environment data standards and their integration: an analysis of IFC , CityGML and LandInfra*. Available at: https://portal.ogc.org/files/?artifact_id=96354.

Gillani, S., Laforest, F. and Picard, G. (2014) 'A generic ontology for prosumer-oriented smart grid', *CEUR Workshop Proceedings*, 1133(c), pp. 134–139.

GIS Standards.EU (2018) *GIS Standards*. Available at: https://www.gistandards.eu/gis-standards/ (Accessed: 9 November 2020).

Giudicianni, C. *et al.* (2018) 'Topological Taxonomy of Water Distribution Networks', *Water (Switzerland)*, 10. doi: 10.3390/w10040444.

Giudicianni, C., Herrera, M., di Nardo, A. and Adeyeye, K. (2020) 'Automatic Multiscale Approach for Water Networks Partitioning into Dynamic District Metered Areas', *Water Resources Management*, 34(2), pp. 835–848. doi: 10.1007/s11269-019-02471-w.

Giudicianni, C., Herrera, M., di Nardo, A., Greco, R., *et al.* (2020) ‘Topological Placement of Quality Sensors in Water-Distribution Networks without the Recourse to Hydraulic Modeling’, *Journal of Water Resources Planning and Management*, 146(6). doi: 10.1061/(ASCE)WR.1943-5452.0001210.

Giustolisi, O., Ridolfi, L. and Simone, A. (2019) ‘Tailoring Centrality Metrics for Water Distribution Networks’, *Water Resources Research*, 55(3), pp. 2348–2369. doi: 10.1029/2018WR023966.

Goldfeather, J. (2015) *CSG Tree, Carleton College*. Available at: http://www.cs.carleton.edu/faculty/jgoldfea/cs311/inclass/csgtree.jpg (Accessed: 12 November 2020).

Goldman, S. L. (2004) ‘Why we need a philosophy of engineering: A work in progress’, *Interdisciplinary Science Reviews*, 29(2), pp. 163–176. doi: 10.1179/030801804225012572.

Grieves, M. W. (2019) *Virtually Intelligent Product Systems: Digital and Physical Twins, Complex Systems Engineering: Theory and Practice*. doi: 10.2514/5.9781624105654.0175.0200.

Guardian News (2018) *Manchester hospital water outage*. Available at: https://www.theguardian.com/uk-news/2018/feb/05/manchester-hospitals-cancel-operations-mains-water-leak (Accessed: 13 November 2021).

Hanus, P., Pęska-Siwik, A. and Szewczyk, R. (2018) ‘Spatial analysis of the accuracy of the cadastral parcel boundaries’, *Computers and Electronics in Agriculture*, 144(November 2017), pp. 9–15. doi: 10.1016/j.compag.2017.11.031.

Hart, W. E. and Murray, R. (2010) ‘Review of Sensor Placement Strategies for Contamination Warning Systems in Drinking Water Distribution Systems’, *Journal of Water Resources Planning and Management*, 136(6), pp. 611–619. doi: 10.1061/(ASCE)WR.1943-5452.0000081.

Hayes, J. and Gutierrez, C. (2004) ‘Bipartite Graphs as Intermediate Model for RDF’, (1030810), pp. 47–61. doi: 10.1007/978-3-540-30475-3_5.

Herle, S. *et al.* (2020) ‘GIM and BIM: How to Obtain Interoperability Between Geospatial and Building Information Modelling?', *PFG - Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88, pp. 33–42. doi: 10.1007/s41064-020-00090-4.

Hetherington, J. and West, M. (2020) *The pathway towards an Information Management Framework - A 'Commons' for Digital Built Britain 30.* Available at: https://www.repository.cam.ac.uk/handle/1810/305579 (Accessed: 24 August 2020).

Hijazi, I. *et al.* (2009) 'IFC to CityGML transformation framework for geo-analysis: a water utility network case', *4th International Workshop on 3D Geo-Information*, pp. 123–127.

Hijazi, I. *et al.* (2011) 'Initial Investigations for Modeling Interior Utilities Within 3D Geo Context: Transforming IFC-Interior Utility to CityGML/UtilityNetworkADE', *Advances in 3D Geo-Information Sciences Lecture Notes in Geoinformation and Cartography*, pp. 73–93. doi: 10.1007/978-3-642-12670-3.

Hijazi, I. (2013) 'Integrated management of indoor and outdoor utilities by utilizing BIM and 3DGIS', p. 171.

HM Government (2013) '2-Construction 2025. Industrial Strategy: Government and industry in partnership', *UK Government*, (July), p. 78. doi: HM Government.

Holtier, S., Steadman, J. P. and Smith, M. G. (2000) 'Three-dimensional representation of urban built form in a GIS', *Environment & Planning B*, 27(1), pp. 51–72. doi: 10.1068/bst5.

Holzschuher, F. and Peinl, R. (2013) 'Performance of graph query languages', *Proceedings of the Joint EDBT/ICDT 2013 Workshops on - EDBT '13*, (March 2013), p. 195. doi: 10.1145/2457317.2457351.

Homayounfar, M. *et al.* (2018) 'Linking resilience and robustness and uncovering their trade-offs in coupled infrastructure systems', *Earth System Dynamics*, 9(4), pp. 1159–1168. doi: 10.5194/esd-9-1159-2018.

Hosseini, S., Barker, K. and Ramirez-Marquez, J. E. (2016) 'A review of definitions and measures of system resilience', *Reliability Engineering and System Safety*, 145, pp. 47–61. doi: 10.1016/j.ress.2015.08.006.

Hu, Z.-Z. and Zhang, J. (2011) 'BIM- and 4D-based integrated solution of analysis and management for conflicts and structural safety problems during construction: 2. Development and site trials', *Automation in Construction*, 20(2), pp. 155–166. doi: 10.1016/j.autcon.2010.09.013.

IBM (2021) *Message brokers*. Available at: https://www.ibm.com/cloud/learn/message-brokers (Accessed: 28 January 2021).

IfcOpenShell contributors (2020) *IfcOpenShell*. Available at: http://ifcopenshell.org/python (Accessed: 28 July 2020).

Institution of Civil Engineers (2014) 'Specification for underground utility detection, verification and location (PAS128:2014)', p. 36.

Irizarry, J. and Karan, E. P. (2012) 'Optimizing Location of Tower Cranes on Construction Sites Through Gis and Bim Integration', *Electronic Journal of Information Technology in Construction*, 17(17), pp. 361–366. Available at: http://www.itcon.org/2012/23.

Irizarry, J., Karan, E. P. and Jalaei, F. (2013) 'Integrating BIM and GIS to improve the visual monitoring of construction supply chain management', *Automation in Construction*. doi: 10.1016/j.autcon.2012.12.005.

Isikdag, U. and Zlatanova, S (2009) 'A SWOT analysis on the implementation of Building Information Models within the Geospatial Environment', *Urban and Regional data Management, UDMS Annuals*, pp. 15–30.

Isikdag, U. and Zlatanova, Sisi (2009) 'Towards Defining a Framework for Automatic Generation of Buildings in CityGML Using Building Information Models', *3D Geo-Information Sciences*, pp. 79–96. doi: 10.1007/978-3-540-87395-2_6.

James, P. *et al.* (2020) 'Urban Observatory Data Newcastle'. doi: https://doi.org/10.25405/data.ncl.c.5059913.v4.

Jaro, T. S. R. D. I. and Iguisi, M. P. E. O. (2015) 'Development of GIS-Based Road Transport Information Management System for Adamawa Central, Adamawa State, Nigeria', *Development*, 5(5), pp. 56–73.

Ji, Q. *et al.* (2017) 'A heuristic spatial algorithm for generating fine-scale infrastructure distribution networks', in *Proceedings of GISRUK*. Available at: http://huckg.is/gisruk2017/GISRUK_2017_paper_37.pdf.

Jiang, B. and Claramunt, C. (2004) ‘Topological analysis of urban street networks’, *Environment and Planning B: Planning and Design*, 31(1), pp. 151–162. doi: 10.1068/b306.

Johansson, J. (2010) *Risk and Vulnerability Analysis of Interdependent Technical Infrastructures: Addressing Socio-Technical Systems*, *Industrial Automation Department of Measurement Technology and Industrial Electrical Engineering, Faculty of Engineering, LTH*. Available at: http://www.iea.lth.se/publications/Theses/LTH-IEA-1061.pdf (Accessed: 20 October 2020).

Johansson, M., Roupé, M. and Bosch-Sijtsema, P. (2015) ‘Real-time visualization of building information models (BIM)’, *Automation in Construction*. doi: 10.1016/j.autcon.2015.03.018.

Jung, D., Lee, S. and Kim, J. H. (2019) ‘Robustness and water distribution system: State-of-the-art review’, *Water (Switzerland)*, 11(5), pp. 1–12. doi: 10.3390/w11050974.

Kamarudin, K. M., Ridgway, K. and Ismail, N. (2016) ‘Abstraction and Generalization in Conceptual Design Process: Involving Safety Principles in TRIZ-SDA Environment’, *Procedia CIRP*, 39, pp. 16–21. doi: 10.1016/j.procir.2016.01.038.

Kang, T. W. and Hong, C. H. (2015) ‘A study on software architecture for effective BIM/GIS-based facility management data integration’, *Automation in Construction*, 54, pp. 25–38. doi: 10.1016/j.autcon.2015.03.019.

Karan, E. P. and Irizarry, J. (2015) ‘Extending BIM interoperability to preconstruction operations using geospatial analyses and semantic web services’, *Automation in Construction*. doi: 10.1016/j.autcon.2015.02.012.

Karan, E. P., Irizarry, J. and Haymaker, J. (2016) ‘BIM and GIS Integration and Interoperability Based on Semantic Web Technology’, *Journal of Computing in Civil Engineering*, 30(3). doi: 10.1061/(ASCE)CP.1943-5487.0000519.

Keary, T. (2020) *Network Topologies Explained & Compared*, *comparitech*. Available at: https://www.comparitech.com/net-admin/network-topologies-advantages-disadvantages/ (Accessed: 12 November 2020).

Kensek, K. M. (2014) *Building Information Modeling*. Pocket Architecture: Technical Design Series.

Khan, W. and Shahzad, W. (2017) 'Predictive Performance Comparison Analysis of Relational & NoSQL Graph Databases', *International Journal of Advanced Computer Science and Applications*, 8(5), pp. 523–530.

Khoa Bui, X., S. Marlim, M. and Kang, D. (2020) 'Water Network Partitioning into District Metered Areas: A State-Of-The-Art Review', *Water*, 12(4). doi: 10.3390/w12041002.

Klein, M. *et al.* (2001) 'The relation between ontologies and XML schemas', *Electronic Trans on AI*, 6(4), pp. 65–94. Available at: http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/etai01.pdf.

KMC Controls (2018) *KMC101: What is BACnet?* Available at: https://www.youtube.com/watch?v=0hX9X0xZ6OE (Accessed: 18 February 2021).

Kolbe, T. H. (2009) 'Representing and Exchanging 3D City Models with CityGML', *3D Geo-Information Sciences*, pp. 15–31. doi: 10.1007/978-3-540-87395-2_2.

Kolbe, T. H. and Groger, G. (2003) 'Toward unified 3D city models', *ISPRS Comm. IV Joint Workshop on 'Challenges in Geospatial Analysis, Integration and Visualization II'*.

Kolbe, T. H., Groger, G. and Plumer, L. (2005) 'CityGML: Interoperable access to 3D city models', *Geo-information for Disaster Management*, pp. 883–899. doi: 10.1007/3-540-27468-5_63.

Kolbe, T. H. and Kutzner, T. (2016) 'Motivation and use cases for 3D utility network models + Utility Network ADE Core Model'.

Koomey, J. and Brown, R. E. (2002) 'The role of building technologies in reducing and controlling peak electricity demand', *Lawrence Berkeley National Laboratory*, (September). Available at: http://escholarship.org/uc/item/3tc7b2kh.pdf.

Kumar, K. *et al.* (2019) 'The LandInfra standard and its role in solving the BIM-GIS quagmire', *Open Geospatial Data, Software and Standards*, 4(1). doi: 10.1186/s40965-019-0065-z.

Kutzner, T. and Kolbe, T. H. (2016) *Current state of development of the CityGML UtilityNetworkADE, 1st Joint SIG 3D and OGC Workshop on the CityGML UtilityNetworkADE Munich, October 13-14, 2016*. Available at:

https://en.wiki.utilitynetworks.sig3d.org/images/upload/20160913_-_Kutzner_-_Current_state_of_development_of_the_CityGML_UtilityNetworkADE.pdf (Accessed: 10 April 2017).

Laakso, M. and Kiviniemi, A. (2012) ‘The IFC standard - A review of history, development, and standardization’, *Electronic Journal of Information Technology in Construction*, 17(May), pp. 134–161.

Laat, R. De and Berlo, L. Van (2011) ‘Integration of BIM and GIS: The development of the CityGML GeoBIM extension’, *Advances in 3D Geo-Information Sciences*, pp. 211–225. doi: 10.1007/978-3-642-12670-3_13.

Lapierre, A. . and Cote, P. . (2007) ‘Using open web services for urban data management: A testbed resulting from an OGC initiative for offering standard CAD/GIS/BIM services’, *Proceedings of the Urban and Regional Data Management - UDMS Annual 2007*, pp. 381–393. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-60749089719&partnerID=40&md5=e8ece4e74b55734aa0c099fc8e023f57.

Lee, J. (2004) ‘A spatial access-oriented implementation of a 3-D GIS topological data model for urban entities’, *GeoInformatica*, 8(3), pp. 237–264. doi: 10.1023/B:GEIN.0000034820.93914.d0.

Lee, K. H. and Braun, J. E. (2008) ‘Evaluation of methods for determining demand-limiting setpoint trajectories in buildings using short-term measurements’, *Building and Environment*, 43(10), pp. 1769–1783. doi: 10.1016/j.buildenv.2007.11.003.

Leon-garcia, A. (2010) ‘Price Prediction in Real-Time Electricity’, *IEEE Transactions on Smart Grid*, 1(2), pp. 120–133. doi: 10.1109/TSG.2010.2055903.

Levy, R. M. (1995) ‘Visualization of urban alternatives’, *Environment & Planning B*, 22(3), pp. 343–358. doi: 10.1068/b220343.

Li, F. *et al.* (2010) ‘Smart Transmission Grid : Vision and Framework’, 1(2), pp. 168–177.

Li, L. *et al.* (2017) ‘Spatial Data Uncertainty’, *Comprehensive Geographic Information Systems*, 3(November), pp. 313–340. doi: 10.1016/B978-0-12-409548-9.09610-X.

Lieberman, J. (2017) *OGC Underground Infrastructure Concept Study Engineering Report*.

Available at: http://www.opengis.net/doc/PER/uicds (Accessed: 4 August 2020).

Lieberman, J. (2018) *Introduction to MUDDI : Model for Underground Data Definition and Integration*. Available at: https://portal.opengeospatial.org/files/?artifact_id=80269 (Accessed: 20 November 2020).

Lieberman, J. *et al.* (2019) *Underground Data Definition and Integration Engineering Report*. Available at: http://docs.opengeospatial.org/per/19-081.pdf (Accessed: 24 August 2020).

Liebich, T. (2010) 'Unveiling IFC2x4 - The next generation of OPENBIM', *Proceedings of CIB W78 Conference*, pp. 124–131.

Liggett, R. S. and Jepson, W. H. (1995) 'An integrated environment for urban simulation', *Environment & Planning B*, 22, pp. 291–302. doi: 10.1068/b220291.

Likhari, R. *et al.* (2017) *Mapping Underground Assets in the UK*. Available at: http://nora.nerc.ac.uk/id/eprint/519019/ (Accessed: 24 August 2020).

Liu, X. *et al.* (2017) 'A State-of-the-Art Review on the Integration of Building Information Modeling (BIM) and Geographic Information System (GIS)', *ISPRS International Journal of Geo-Information*, 6(2). doi: 10.3390/ijgi6020053.

Longley, P. A. *et al.* (2011) *Geographic Information Systems & Science*.

Ma, J. *et al.* (2016) 'Residential Load Scheduling in Smart Grid: A Cost Efficiency Perspective', *IEEE Transactions on Smart Grid*, 7(2), pp. 771–784. doi: 10.1109/TSG.2015.2419818.

Marlow, P. (2005) *Boundary Representation*. Available at: https://paul-marlow.squarespace.com/blog/2020/6/25/geometric-abstraction-exploded-cube-summer-solstice (Accessed: 24 September 2020).

Martinelli, A. (2012) 'An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry', *Research Policy*, 41(2), pp. 414–429. doi: 10.1016/j.respol.2011.10.012.

Matthews, J. C. (2016) 'Disaster Resilience of Critical Water Infrastructure Systems', *Journal* *of Structural Engineering (United States)*, 142(8). doi: 10.1061/(ASCE)ST.1943-541X.0001341.

Meng, F. *et al.* (2018) 'Topological attributes of network resilience: A study in water distribution systems', *Water Research*, 143, pp. 376–386. doi: 10.1016/j.watres.2018.06.048.

El Meouche, R., Rezoug, M. and Hijazi, I. (2013) 'Integrating and managing BIM in gis, software review', *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, XL-2/W2(November), pp. 31–34. doi: 10.5194/isprsarchives-XL-2-W2-31-2013.

Metje, N. *et al.* (2020) 'Improved underground utilities asset management-Assessing the impact of the UK utility survey standard (PAS128)', *Proceedings of the Institution of Civil Engineers: Municipal Engineer*, 173(4), pp. 218–236. doi: 10.1680/jmuen.18.00055.

Metje, N., Ahmad, B. and Crossland, S. M. (2015) 'Causes, impacts and costs of strikes on buried utility assets', *Proceedings of the Institution of Civil Engineers: Municipal Engineer*, 168(ME3), pp. 165–174. doi: 10.1680/muen.14.00035.

Mignard, C. and Nicolle, C. (2014) 'Merging BIM and GIS using ontologies application to Urban facility management in ACTIVe3D', *Computers in Industry*. doi: 10.1016/j.compind.2014.07.008.

Mirshafiei, P. *et al.* (2019) 'Geospatial information system-based modeling approach for leakage management in urban water distribution networks', *Water (Switzerland)*, 11(8). doi: 10.3390/w11081736.

MISTRAL (2020) *Interdependencies in infrastructure systems-of-systems*. Available at: https://www.itrc.org.uk/highlights/interdependencies-in-infrastructure-systems-of-systems/ (Accessed: 11 August 2020).

Moholkar, A., Klinkhachorn, P. and Feliachi, A. (2004) 'Effects of dynamic pricing on residential electricity bill', *2004 IEEE PES Power Systems Conference and Exposition*, 2, pp. 1030–1035. doi: 10.1109/psce.2004.1397738.

Mohsenian-Rad, A. H. *et al.* (2010) 'Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid', *IEEE Transactions on*

Smart Grid, 1(3), pp. 320–331. doi: 10.1109/TSG.2010.2089069.

Nagarajan, H. (2018) 'On Maximizing Weighted Algebraic Connectivity for Synthesizing Robust Networks'. Available at: http://arxiv.org/abs/1805.07825.

di Nardo, A. *et al.* (2018) 'Applications of graph spectral techniques towater distribution network management', *Water (Switzerland)*, 10(1), pp. 1–16. doi: 10.3390/w10010045.

Di Nardo, A., Di Natale, M., Giudicianni, C., Musmarra, D., *et al.* (2017) 'Redundancy features of water distribution systems', *Procedia Engineering*, 186, pp. 412–419. doi: 10.1016/j.proeng.2017.03.244.

Di Nardo, A., Di Natale, M., Giudicianni, C., Greco, R., *et al.* (2017) 'Weighted spectral clustering for water distribution network partitioning', *Applied Network Science*, 2(1). doi: 10.1007/s41109-017-0033-4.

National Infrstructure Commission (2017) 'Data for the Public Good', p. 23.

National Institute of Building Sciences (2007) 'National Building Information Modeling Standard', *Nist*. Available at: http://academics.triton.edu/faculty/fheitzman/NBIMSv1_ConsolidatedBody_11Mar07_4.pdf.

NBS (2011) *NBS: What is COBie?* Available at: https://www.thenbs.com/knowledge/what-is-cobie (Accessed: 1 August 2017).

NBS (2017) *NBS: BIM Levels explained*. Available at: https://www.thenbs.com/knowledge/bim-levels-explained (Accessed: 9 January 2017).

NBS (2021) *Exploring the NBS BIM Object Standard*. Available at: https://www.thenbs.com/knowledge/exploring-the-nbs-bim-object-standard (Accessed: 3 March 2020).

NCL University (2017) *Newcastle Urban Observatory*. Available at: http://uoweb1.ncl.ac.uk/ (Accessed: 5 January 2017).

Neo4j (2021) *Neo4j Graph Database*. Available at: https://neo4j.com/product/neo4j-graph-database/ (Accessed: 3 March 2021).

Newcastle University (2018a) *Newcastle Helix*. Available at:

https://newcastlehelix.com/about (Accessed: 8 December 2020).

Newcastle University (2018b) *Urban Sciences Building - Description*. Available at: http://www.ncl.ac.uk/sciencecentral/urban/ (Accessed: 8 March 2018).

Newcastle University (2018c) *Urban Sciences Building Brochure*. Available at: http://www.ncl.ac.uk/sciencecentral/urban/.

Newcastle University (2020) *Frederick Douglass Centre*. Available at: https://newcastlehelix.com/about/frederick-douglass-centre (Accessed: 10 December 2020).

Newcastle University (2021) *Catalyst*. Available at: https://www.thecatalystnewcastle.co.uk/ (Accessed: 9 March 2021).

Nisbet, N. (2015) 'Using Linked Open Data'. AEC3.

Niu, S., Pan, W. and Zhao, Y. (2015) 'A BIM-GIS Integrated Web-based Visualization System for Low Energy Building Design', in *Procedia Engineering*. doi: 10.1016/j.proeng.2015.09.091.

Noardo, F. *et al.* (2020) 'Tools for BIM-GIS Integration (IFC Georeferencing and Conversions): Results from the GeoBIM Benchmark 2019', *ISPRS International Journal of Geo-Information*, 9(9), p. 502. doi: 10.3390/ijgi9090502.

Northumbrian Water (2018) *SELF LAY LOCAL REQUIREMENTS*. Available at: https://www.nwl.co.uk/globalassets/customer-pdfs/developer-pdfs/draft_self_lay_document_final.pdf.

Northumbrian Water Group (2019) 'Living water: Our plan 2020-2025 and beyond', pp. 1–305.

Ofwat (2015) *Towards Water 2020 – meeting the challenges for water and wastewater services in England and Wales*. Available at: https://www.ofwat.gov.uk/wp-content/uploads/2015/10/pap_tec201507challenges.pdf (Accessed: 29 May 2020).

OGC (2012) *OGC CityGML Standard, OGC Member approved international standard*. doi: OGC 12-019.

OGC (2014) *OGC IndoorGML Standard, OGC Member approved international standard*.

doi: http://www.opengeospatial.org/.

OGC (2016) ‘OGC® Land and Infrastructure Conceptual Model Standard (LandInfra)’.

OGC (2019) ‘OGC Abstract Specification Topic 2: Referencing by coordinates’, p. 150. Available at: http://docs.opengeospatial.org/as/18-005r4/18-005r4.html.

OGC (2020) *IndoorGML*. Available at: http://www.indoorgml.net/ (Accessed: 20 November 2020).

Ohori, K. A. *et al.* (2018) ‘Processing BIM and GIS models in practice: Experiences and recommendations from a GeoBIM project in The Netherlands’, *ISPRS International Journal of Geo-Information*, 7(8). doi: 10.3390/ijgi7080311.

Oliker, N. and Ostfeld, A. (2013) ‘Multi-objective optimization of cost and resilience of water distribution system design’, *World Environmental and Water Resources Congress 2013: Showcasing the Future - Proceedings of the 2013 Congress*, pp. 845–856. doi: 10.1061/9780784412947.081.

omni.sci (2020) *Data Integration*, *omni.sci*. Available at: https://www.omnisci.com/technical-glossary/data-integration (Accessed: 12 November 2020).

OmniClass (2017) *OmniClass*. Available at: http://www.omniclass.org/about/ (Accessed: 1 July 2017).

Open Geospatial Consortium (2015) ‘OGC Smart Cities Spatial Information Framework’, pp. 1–32. doi: http://www.opengeospatial.org/.

Open Geospatial Consortium (2017) *The OGC Community Standard Process*. Available at: https://www.ogc.org/blog/2543 (Accessed: 23 February 2021).

Ordnance Survey (2013) *Linked Data Blog*, *OS blog*. Available at: https://www.ordnancesurvey.co.uk/blog/2013/06/new-linked-data-service-launches/#more-10434 (Accessed: 18 November 2020).

Ordnance Survey (2020a) *Linked Data Service*, *OS web service*. Available at: http://data.ordnancesurvey.co.uk/datasets/os-linked-data (Accessed: 18 November 2020).

Ordnance Survey (2020b) ‘Positional Accuracy - Absolute Accuracy’.

Pauwels, P., Zhang, S. and Lee, Y. C. (2017) ‘Semantic web technologies in AEC industry: A literature overview’, *Automation in Construction*, 73, pp. 145–165. doi: 10.1016/j.autcon.2016.10.003.

Peachavanish, R. *et al.* (2006) ‘An ontological engineering approach for integrating CAD and GIS in support of infrastructure management’, *Advanced Engineering Informatics*, 20(1). doi: 10.1016/j.aei.2005.06.001.

Peckiene, A. and Ustinovičius, L. (2017) ‘Possibilities for Building Spatial Planning using BIM Methodology’, *Procedia Engineering*, 172, pp. 851–858. doi: 10.1016/j.proeng.2017.02.085.

Perelman, L. and Ostfeld, A. (2011) ‘Topological clustering for water distribution systems analysis’, *Environmental Modelling and Software*, 26(7), pp. 969–972. doi: 10.1016/j.envsoft.2011.01.006.

Peters, E. (2010) ‘BIM and Geospatial Information Systems’, *Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies*, (2003), pp. 483–500. doi: 10.4018/978-1-60566-928-1.ch021.

Ponsen, M., Taylor, M. E. and Tuyls, K. (2010) *Abstraction and generalization in reinforcement learning: A summary and framework*, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-11814-2_1.

Portele, C. *et al.* (2007) ‘Geography Markup Language (GML) Encoding Standard’, *OGC Implementation Specification*, 07–036, p. viii + 426. Available at: http://portal.opengeospatial.org/files/20509.

Pritchard, O. G., Hallett, S. H. and Farewell, T. S. (2013) ‘Infrastructure Transitions Research Consortium Working paper series Soil movement in the UK-Impacts on critical infrastructure Soil movement in the UK-Impacts on Critical Infrastructure’. Available at: https://www.itrc.org.uk/wp-content/PDFs/Soil-movement-impacts-UK-infrastructure.pdf.

Propertymark (2021) ‘Transforming the residential property market: Driving adoption of the UPRN across the sector’. Available at: https://www.arla.co.uk/news/january-2021/property-industry-urges-the-use-of-the-uprn-across-the-sector.aspx.

Qadrdan, M. *et al.* (2017) 'Benefits of demand-side response in combined gas and electricity networks', *Applied Energy*, 192, pp. 360–369. doi: 10.1016/j.apenergy.2016.10.047.

Rafiee, A. *et al.* (2014) 'From BIM to Geo-analysis: View Coverage and Shadow Analysis by BIM/GIS Integration', *Procedia Environmental Sciences*, 22(0), pp. 397–402. doi: 10.1016/j.proenv.2014.11.037.

Reed, C. and OGC (2018) 'Open Geospatial Consortium Volume 8 : CDB Spatial and Coordinate Reference Systems Guidance', pp. 1–19.

Resch, B. *et al.* (2014) 'GIS-Based Planning and Modeling for Renewable Energy: Challenges and Future Research Avenues', *ISPRS International Journal of Geo-Information*, 3(2), pp. 662–692. doi: 10.3390/ijgi3020662.

Rezaei, P. (2016) 'Cascading failure risk estimation and mitigation in power systems', *ProQuest Dissertations and Theses*, p. 117. Available at: http://ezproxy.rice.edu/login?url=https://search.proquest.com/docview/1772350009?accountid=7064%0Ahttp://sfxhosted.exlibrisgroup.com/rice?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ProQuest+Di.

Rizal, S., Michel, B. and Pim Van Den, H. (2013) 'BIM and GIS for low-disturbance construction', *Proceedings of the 13th International Conference on Construction Applications of Virtual Reality*, (October), pp. 30–31. Available at: http://itc.scix.net/data/works/att/convr-2013-47.pdf.

Robson, C. (2016) 'Robustness of Hierarchical Spatial Critical Infrastructure Networks', (June), p. 221.

Romero Rodríguez, L. *et al.* (2017) 'Assessment of the photovoltaic potential at urban level based on 3D city models: A case study and new methodological approach', *Solar Energy*, 146, pp. 264–275. doi: 10.1016/j.solener.2017.02.043.

Rüppel, U., Abolghasemzadeh, P. and Stübbe, K. (2010) 'BIM-based immersive indoor graph networks for emergency situations in buildings', *International Conference on Computing in Civil and Building Engineering*, (August 2006), pp. 1–7.

Saleh, M., Esa, Y. and Mohamed, A. (2018) ‘Applications of complex network analysis in electric power systems’, *Energies*, 11(6). doi: 10.3390/en11061381.

Saleh, S. A., Pijnenburg, P. and Castillo-Guerra, E. (2017) ‘Load Aggregation from Generation-Follows-Load to Load-Follows-Generation: Residential Loads’, *IEEE Transactions on Industry Applications*, 53(2), pp. 833–842. doi: 10.1109/TIA.2016.2626261.

Sánchez-De-Rivera, D. *et al.* (2017) ‘Adaptation of ontology sets for water related scenarios management with IoT systems for a more productive and sustainable agriculture systems’, *CEUR Workshop Proceedings*, 2063.

Sani, M. J. and Rahman, A. A. (2018) ‘GIS and BIM integration at data level: A review’, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(4/W9), pp. 299–306. doi: 10.5194/isprs-archives-XLII-4-W9-299-2018.

Scarpa, F., Lobba, A. and Becciu, G. (2016) ‘Elementary DMA Design of Looped Water Distribution Networks with Multiple Sources’, *Journal of Water Resources Planning and Management*, 142(6). doi: 10.1061/(ASCE)WR.1943-5452.0000639.

Schneider, M. and Behr, T. (2006) ‘Topological relationships between complex spatial objects’, *ACM Transactions on Database Systems*, 31(1), pp. 39–81. doi: 10.1145/1132863.1132865.

Schürr, A. (1995) ‘Specification of graph translators with triple graph grammars’, pp. 151–163. doi: 10.1007/3-540-59071-4_45.

sensorsandsystems.com (2008) *Interview with Thomas Kolbe*, *V1 Magazine*. Available at: http://sensorsandsystems.com/interview-citygml-modeling-the-city-for-the-future/ (Accessed: 18 September 2017).

SERL (2021) *SERL*, *UCL*. Available at: https://www.ucl.ac.uk/bartlett/energy/research/research-projects-and-centres/smart-energy-research-lab-serl/serl-public-notice-data (Accessed: 19 February 2021).

Shackelford, T. (2017) *The History and Furure of Revit MEP*. Available at: https://www.augi.com/articles/detail/the-history-and-future-of-revit-mep (Accessed: 20

November 2020).

Shahinmoghadam, M. and Motamedi, A. (2019) 'Review of BIM-centred IoT deployment: State of the art, opportunities, and challenges', *Proceedings of the 36th International Symposium on Automation and Robotics in Construction, ISARC 2019*, (Isarc), pp. 1268–1275. doi: 10.22260/isarc2019/0170.

Shi, J., Wang, Y. and Ng, C. W. W. (2013) 'Buried pipeline responses to ground displacements induced by adjacent static pipe bursting', *Canadian Geotechnical Journal*, 50(5), pp. 481–492. doi: 10.1139/cgj-2012-0304.

Siano, P. (2014) 'Demand response and smart grids - A survey', *Renewable and Sustainable Energy Reviews*, 30, pp. 461–478. doi: 10.1016/j.rser.2013.10.022.

Simone, A. *et al.* (2018) 'Complex Network Theory for Water Distribution Networks Analysis', 3(November), pp. 1971–1962. doi: 10.29007/w1bk.

Skandhakumar, N. *et al.* (2016) 'Graph theory based representation of building information models for access control applications', *Automation in Construction*. doi: 10.1016/j.autcon.2016.04.001.

Solomakhina, N. *et al.* (2015) 'Modeling and analysis techniques for multimodal utility networks', *Energy Procedia*, 78, pp. 3397–3402. doi: 10.1016/j.egypro.2015.11.757.

Song, Y. *et al.* (2017) 'Trends and Opportunities of BIM-GIS Integration in the Architecture, Engineering and Construction Industry: A Review from a Spatio-Temporal Statistical Perspective', *ISPRS International Journal of Geo-Information*, 6(12). doi: 10.3390/ijgi6120397.

Srinivasan, D. *et al.* (2017) 'Game-Theory based dynamic pricing strategies for demand side management in smart grids', *Energy*, 126, pp. 132–143. doi: 10.1016/j.energy.2016.11.142.

Stadler, A. and Kolbe, T. H. (2007) 'Spatio-semantic coherence in the integration of 3D city models', *Proceedings of the 5th International ISPRS Symposium on Spatial Data Quality ISSDQ 2007 in Enschede, The Netherlands, 13-15 June 2007*. Available at: http://www.isprs.org/proceedings/XXXVI/2-C43/Session1/paper_Stadler.pdf.

Stouffs, R., Tauscher, H. and Biljecki, F. (2018) 'Achieving Complete and Near-Lossless

Conversion from IFC to CityGML', *ISPRS International Journal of Geo-Information*, 7(9), p. 355. doi: 10.3390/ijgi7090355.

Sun, Y., Wang, S. and Huang, G. (2010) 'A demand limiting strategy for maximizing monthly cost savings of commercial buildings', *Energy and Buildings*, 42(11), pp. 2219–2230. doi: 10.1016/j.enbuild.2010.07.018.

Taher, S. A. and Labadie, J. W. (1996) 'OPTIMAL DESIGN OF WATER-DISTRIBUTION NETWORKS WITH By Saud A. Taher/ and John W. Labadie,z Member, ASCE', 122(August), pp. 301–311.

Taillefond, N. and Wolkenhauer, O. (2002) 'Fuzzy clustering and classification for automated leak detection systems', *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 35(1), pp. 407–411. doi: 10.3182/20020721-6-es-1901.00799.

talend (2020) *What is Data Integration?* Available at: https://www.talend.com/resources/what-is-data-integration/ (Accessed: 20 November 2020).

Tang, K., Parsons, D. J. and Jude, S. (2019) 'Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system', *Reliability Engineering and System Safety*, 186, pp. 24–36. doi: 10.1016/j.ress.2019.02.001.

Tang, R., Wang, S. and Li, H. (2019) 'Game theory based interactive demand side management responding to dynamic pricing in price-based demand response of smart grids', *Applied Energy*, 250(May), pp. 118–130. doi: 10.1016/j.apenergy.2019.04.177.

Tang, R., Wang, S. and Shan, K. (2018) 'Optimal and near-optimal indoor temperature and humidity controls for direct load control and proactive building demand response towards smart grids', *Automation in Construction*, 96(March), pp. 250–261. doi: 10.1016/j.autcon.2018.09.020.

Thames Water (2015) *Mains design guidelines for self-lay organisations working in new developments*. Available at: https://www.thameswater.co.uk/-/media/Site-Content/Developer-Services/Mains-design-guidelines.pdf (Accessed: 27 April 2020).

Thames Water (2021) *Getting a smart water meter*. Available at: https://www.thameswater.co.uk/help/water-meters/getting-a-water-meter (Accessed: 8 March

2021).

Tobiáš, P. (2015) ‘An Investigation into the Possibilities of BIM and GIS Cooperation and Utilization of GIS in the BIM Process’, *Geoinformatics FCE CTU*, 14(1), p. 65. doi: 10.14311/gi.14.1.5.

Torres, J. M. *et al.* (2017) ‘Exploring Topological Effects on Water Distribution System Performance Using Graph Theory and Statistical Models’, *Journal of Water Resources Planning and Management*, 143(1). doi: 10.1061/(ASCE)WR.1943-5452.0000709.

Uggla, G. and Horemuz, M. (2018) ‘Geographic capabilities and limitations of Industry Foundation Classes’, *Automation in Construction*, 96(March), pp. 554–566. doi: 10.1016/j.autcon.2018.10.014.

UK BIM Alliance (2019) ‘Information Management according to BS EN ISO 19650 - Guidance Part 1: Concepts’, *UK BIM Alliance*, (April), pp. 1–42.

UK BIS (2013) *SMART CITIES: Background paper*.

UK Infrastructure and Projects Authority (2016) *Government Construction Strategy 2016-20*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/510354/Government_Construction_Strategy_2016-20.pdf.

UK Parliamentary Office of Science & Technology (2014) *Electricity Demand-Side Response*, *POSTnote*. Available at: http://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-452.

UK Power Networks (2018) ‘MPAN Guide’, pp. 1–3.

US Environmental Protection Agency (2020) *Epanet*. Available at: https://www.epa.gov/water-research/epanet (Accessed: 16 February 2021).

UtilityNetwork ADE contributors (2011) ‘UtilityNetwork ADE Core Schema’, pp. 1–27.

Vishnu, E. and Saran, S. (2018) ‘Semantic modeling of utility networks implementation of use cases for Dehradun city’, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(5), pp. 139–145. doi: 10.5194/isprsarchives-XLII-5-139-2018.

Visual Paradigm (2021) *UML Association vs Aggregation vs Composition*. Available at: https://www.visual-paradigm.com/guide/uml-unified-modeling-language/uml-aggregation-vs-composition/ (Accessed: 18 September 2021).

Volk, R., Stengel, J. and Schultmann, F. (2014) 'Building Information Modeling (BIM) for existing buildings - Literature review and future needs', *Automation in Construction*, 38, pp. 109–127. doi: 10.1016/j.autcon.2013.10.023.

W3 (2014) *RDF 1.1 Primer*. Available at: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/ (Accessed: 26 October 2020).

W3C (2011a) *Notation3 (N3): A readable RDF syntax*. Available at: https://www.w3.org/TeamSubmission/n3/ (Accessed: 26 October 2020).

W3C (2011b) *Turtle - Terse RDF Triple Language*. Available at: https://www.w3.org/TeamSubmission/turtle/ (Accessed: 26 October 2020).

W3C (2014) *RDF Schema 1.1*. Available at: https://www.w3.org/TR/rdf-schema/ (Accessed: 26 October 2020).

Wang, H., Pan, Y. and Luo, X. (2019) 'Integration of BIM and GIS in sustainable built environment: A review and bibliometric analysis', *Automation in Construction*, 103, pp. 41–52. doi: 10.1016/j.autcon.2019.03.005.

Wang, X. (2009) 'Augmented Reality in Architecture and Design: Potentials and Challenges for Application', *International Journal of Architectural Computing*, 7(2), pp. 309–326. doi: 10.1260/147807709788921985.

Wang, Y. *et al.* (2019) 'Ground deformation analysis using InSAR and backpropagation prediction with influencing factors in Erhai Region, China', *Sustainability (Switzerland)*, 11(10). doi: 10.3390/su11102853.

Water UK (2018) 'Sewers for Adoption', (September), p. 48. Available at: http://sfa.wrcplc.co.uk/Data/Sites/4/GalleryImages/WebImages/pdfs/SFA7smalldev_sep2013.pdf.

WBDG (2017) *Whole Building Design Guide*. Available at: https://www.wbdg.org/resources/construction-operations-building-information-exchange-cobie (Accessed: 1 August 2017).

Webborn, E. *et al.* (2019) 'Utilising smart meter data for research and innovation in the UK', *Eceee Summer Study Proceedings*, 2019-June, pp. 1387–1396.

Wei, P. and Sun, D. (2011) *Weighted algebraic connectivity: An application to airport transportation network*, *IFAC Proceedings Volumes (IFAC-PapersOnline)*. IFAC. doi: 10.3182/20110828-6-IT-1002.00486.

Weygant, R. S. (2011) *BIM content development standards, strategies, and best practices*. John Wiley & Sons.

Whiting, E. J. and Teller, S. (2006) 'Geometric, Topological & Semantic Analysis of Multi-Building Floor Plan Data', *Department of Architecture*, Master Of. Available at: http://rvsn.csail.mit.edu/Pubs/master_whiting_2006june_bmgwriteup.pdf.

Wilkinson, M. D. *et al.* (2016) 'Comment: The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 3, pp. 1–9. doi: 10.1038/sdata.2016.18.

Williams, V. V. (2012) 'Multiplying matrices faster than Coppersmith-Winograd', *Proceedings of the Annual ACM Symposium on Theory of Computing*, pp. 887–898. doi: 10.1145/2213977.2214056.

Wilson, J. P. and Fotheringham, A. S. (2008) *The Handbook of Geographic Information Science*.

Wright, R. *et al.* (2014) 'Adaptive water distribution networks with dynamically reconfigurable topology', *Journal of Hydroinformatics*, 16(6), pp. 1280–1301. doi: 10.2166/hydro.2014.086.

Wright, R. *et al.* (2015) 'Optimized control of pressure reducing valves in water distribution networks with dynamic topology', *Procedia Engineering*, 119(1), pp. 1003–1011. doi: 10.1016/j.proeng.2015.08.994.

Wrozynski, R., Sojka, M. and Pyszny, K. (2016) 'The application of GIS and 3D graphic software to visual impact assessment of wind turbines'. doi: 10.1016/j.renene.2016.05.016.

Wu, B. and Zhang, S. (2016) ‘Integration of GIS And BIM for indoor geovisual analytics’, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July), pp. 455–458. doi: 10.5194/isprsarchives-XLI-B2-455-2016.

Wu, I. and Hsieh, S. (2007) ‘Transformation from IFC data model to GML data model: Methodology and tool development’, *Journal of the Chinese Institute of Engineers*, 30(6), pp. 1085–1090. doi: 10.1080/02533839.2007.9671335.

Wu, Y. *et al.* (2016) ‘Burst detection in district metering areas using a data driven clustering algorithm’, *Water Research*, 100, pp. 28–37. doi: 10.1016/j.watres.2016.05.016.

Xia, Y. (2010) ‘Synthetic aperture radar interferometry’, *Sciences of Geodesy - I: Advances and Future Directions*, 88(3), pp. 415–474. doi: 10.1007/978-3-642-11741-1_11.

Yan, J., Van Son, R. and Soon, K. H. (2021) ‘From underground utility survey to land administration: An underground utility 3D data model’, *Land Use Policy*, 102, p. 105267. doi: 10.1016/j.landusepol.2020.105267.

Yao, Z. *et al.* (2018) ‘3DCityDB - a 3D geodatabase solution for the management, analysis, and visualization of semantic 3D city models based on CityGML’, *Open Geospatial Data, Software and Standards*, 3(1). doi: 10.1186/s40965-018-0046-7.

Yazdani, A. and Jeffrey, P. (2010) ‘A complex network approach to robustness and vulnerability of spatially organized water distribution networks’, pp. 1–18. Available at: http://arxiv.org/abs/1008.1770.

Yazdani, A. and Jeffrey, P. (2011) ‘Complex network analysis of water distribution systems’, *Chaos*, 21(1). doi: 10.1063/1.3540339.

Yazdani, A. and Jeffrey, P. (2012) ‘Applying Network Theory to Quantify the Redundancy and Structural Robustness of Water Distribution Systems’, *Journal of Water Resources Planning and Management*, 138(2), pp. 153–161. doi: 10.1061/(asce)wr.1943-5452.0000159.

Yorkshire Water (2018) *Mains Design and Construction Guidelines for Self-Lay Providers*. Available at: https://www.yorkshirewater.com/media/xfvntxqd/mains-design-and-construction-guidelines-for-self-lay-providers-v3-jan-2020-1.pdf.

Zhang, X. *et al.* (2009) 'Integrating BIM and GIS for large scale (building) asset management : a critical review', *Management*, (March), pp. 1–15.

Zhao, L., Liu, Z. and Mbachu, J. (2019) 'An Integrated BIM–GIS Method for Planning of Water Distribution System', *ISPRS International Journal of Geo-Information,* 8. doi: 10.3390/ijgi8080331.

Zheng, Y. *et al.* (2017) 'Weighted Algebraic Connectivity Maximization for Optical Satellite Networks', *IEEE Access*, 5, pp. 6885–6893. doi: 10.1109/ACCESS.2017.2697818.

Zhu, J. *et al.* (2019) 'Integration of BIM and GIS: Geometry from IFC to shapefile using open-source technology', *Automation in Construction*, 102(February), pp. 105–119. doi: 10.1016/j.autcon.2019.02.014.

Žiha, K. (2000) 'Redundancy and robustness of systems of events', *Probabilistic Engineering Mechanics*, 15(4), pp. 347–357. doi: 10.1016/S0266-8920(99)00036-3.

Zirak, M., Royapoor, M. and Gilbert, T. (2019) 'Cross-platform Energy Modeling for Scalable Urban Energy Simulation : a Case-Study', pp. 2–7.