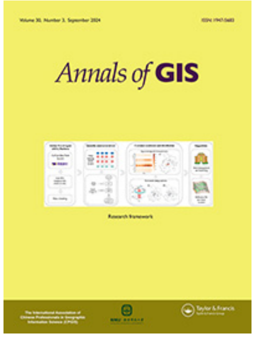

Annals of GIS

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tagi20

Data integration across urban digital twin lifecycle: a comprehensive review of current initiatives

Imane Jeddoub, Gilles-Antoine Nys, Rafika Hajji & Roland Billen

ANNALS OF GIS
https://doi.org/10.1080/19475683.2024.2416135

REVIEW

OPEN ACCESS Check for updates

Data integration across urban digital twin lifecycle: a comprehensive review of current initiatives

Imane Jeddoub[a], Gilles-Antoine Nys[a], Rafika Hajji[b] and Roland Billen[a]

[a]GeoScITY, UR SPHERES, University of Liège, Liège, Belgium; [b]College of Geomatic Sciences and Surveying Engineering, Hassan II Institute of Agronomy and Veterinary Medicine, Rabat, Morocco

ABSTRACT

Challenges related to data integration and interoperability were raised recently under the auspice of the Urban Digital Twin (UDT). This new paradigm shows its potential to address current city challenges. However, to maximize its outcomes at the city scale, we should tackle the fundamental issues related to data integration. Indeed, various Digital Twin (DT) frameworks are developed in practice. Their implementations led to the identification of three main levels of data integration. The first level involves the extension of the data model to handle new information. The second level supports data by default, and the data needs to be transformed to meet the model requirements. The third level performs the integration at the front-end level with the help of system architectures. The aim of this work is to analyze, illustrate, and guide the effectiveness of different data integration approaches. This exploratory review unpacks the levels of integration according to the corresponding UDT lifecycle phases (i.e., creation, use, and update phases). It highlights the challenges and potentialities of data integration levels and offers the DT designer conceptual guidelines related to data integration. Furthermore, current and theoretical data integration scenarios are extracted and investigated, considering several types and sources of data. This research provides a comprehensive analytical framework for data integration within UDTs, where some of the current operational UDT are examined based on the various integration levels of life cycle data. While the state-of-the-art identifies data integration as a major challenge for the full implementation of UDT, it is not explored in depth, and the integration is only addressed from a case study-specific perspective, according to the data availability and the UDT requirement. Hence, this framework provides a generic and urban application-independent overview of the different levels of data integration based on the UDT lifecycle inspired by the Spatial Data Infrastructure lifecycle. This article provides first conceptual insights of data integration levels to build, use, and update UDT. However, from a practical perspective, the list of UDT initiatives used to illustrate the work is not exhaustive, and future initiatives should be documented. Furthermore, the current emphasis is on the creation and use phases of the lifecycle, which lacks a concrete illustration of the update phase. Indeed, it limits the practicability of the data integration levels in the maintenance phase.

ARTICLE HISTORY
Received 28 June 2024
Accepted 8 October 2024

KEYWORDS
Urban digital twins; data integration; level of integration; data interoperability; DT lifecycle

1. Introduction

Urban Digital Twins (UDTs) are fast establishing a foothold as a contemporary trend and a research hotspot worldwide in the urban and geospatial fields. They have triggered a huge interest in the current technological discourse. Nevertheless, this new technology comes with its own challenges, restricting its full implementation at the city level. For instance, data integration has significantly triggered considerable interest as one of the severe challenges while moving towards the implementation of UDTs. Although integrating the existing data might address the incompleteness of the datasets, minimize the costs of new acquisitions, and create new and insightful data flows towards more advanced use cases and urban applications. However, the reuse and integration of available and accessible data raise data interoperability and integration issues (Lei et al. 2023; Noardo 2022). Furthermore, recent studies leverage novel data integration paradigms and challenges of Generative AI models and Urban Digital Twins under the Cognitive Digital Twin concept (Xu et al. 2024). This Artificial intelligence (AI)-driven Digital Twin is an emerging form of the traditional DT that is augmented with cognitive functions and extended with semantic technologies. That means that Cognitive Digital Twins is a sophisticated and enhanced stage of the DT implementation. They reveal a promising evolution of the current DT concept by enhancing DT capabilities towards more intelligent, autonomous, self-learning, and reasoning twins. This enables the digital counterpart

CONTACT Imane Jeddoub i.jeddoub@uliege.be Faculty of Science - Department of Geography - UR SPHERES - Geospatial Data Science and City Information Modelling (GeoScITY), University of Liège, Bât. B5A, Quartier Agora allée du Six août, Liège 19 4000, Belgique

to learn from the provided data to address the data complexity based on the deep learning approaches, simulate and predict complex scenarios, and improve autonomous and real-time decision-making. The particularity of this cognitive digital twin is the capacity to continuously evolve according to the physical world across the entire lifecycle (Zheng, Lu, and Kiritsis 2022).

To highlight the close relationship between data integration and UDT implementation, it is first important to clarify how Digital Twins (DTs) are conceived within the scope of this work. DTs are defined as digital models of the real world, allowing data exchange between these two counterparts. Since the concept has been used interchangeably in a transdisciplinary way according to different backgrounds (Kleftakis et al. 2022), we adopt the following definition from an urban and geospatial perspective: the urban or geospatial DTs use 3D city models (3DCM) enriched with semantic information (i.e. enhancing the potential of geospatial data), often coupled with near real-time data (emphasizing the dynamic aspect of DTs), enabling data flow between the real and virtual worlds, offering various sets of analysis through simulations, predictions, and visualization tools (web-based applications, analytical tools or game engines platforms), and creating a one-stop platform (i.e. a multiscale and multitemporal database) of various datasets and models enabling them to investigate and address city current and future social, economic, and environmental challenges, to name a few (Alva, Biljecki, and Stouffs 2022; Jeddoub et al. 2023; Ketzler et al. 2020; Stoter, Arroyo Ohori, and Noardo 2021).

Following the mapping of the challenges with the DT lifecycle, data integration is commonly considered in the processing and generating phases (Lei et al. 2023). However, the use of DTs – more precisely, the use of 3DCM (as integral components of UDTs) and as input data for simulations and urban analysis – generates, in turn, new and specific datasets according to the use case that need to be integrated into UDTs. In addition, the maintenance phase (i.e. the update) requires updated data as it becomes available to be ingested and processed. This implies that integration is part of not only the processing and generation phases of the DT lifecycle but also extends to the stages of use, maintenance, and updating.

Current implementations reveal a lack of a specific and generic approach to data integration. This is justified by the fact that, on the one hand, data integration is carried out through various approaches, as will be discussed in the following paper. On the other hand, for each specific case, the data to be processed varies, and so does the integration approach. Moreover, this integration issue has always been encountered in related and similar concepts (3DCM, spatial data infrastructure, city information models, etc.) for more than a decade, yet there are still considerable efforts being conducted to achieve seamless integration and enhance data interoperability. Thus, data integration needs to be tackled in a comprehensive way, especially while using 3DCMs as the integration basis for UDTs. Using standardized models (i.e. CityGML as the common data model for semantic 3DCMs and recently UDTs) shows its potential to solve some integration issues. For instance, the enrichment and integration of external data into the 3DCMs are feasible through application domain extensions (ADEs) according to the use case. However, as many cities move towards the implementation of UDTs, integration of different data sets (i.e. Internet of Things (IoT) data, simulation outputs, external databases, BIM (Building Information Model), etc.) with different formats, sources, and data quality is still considered one of the first issues in practice.

To support data integration, we need to understand how data might be integrated into UDTs in practice according to the corresponding DT lifecycle stages. This lifecycle, which was borrowed from the spatial data infrastructure stages, mainly consists of three main phases: creation, use, and maintenance. The purpose of establishing such a link between data integration and the different lifecycle phases is to provide fundamental insights on the underlying questions: How do we incorporate data to create a DT, which may be a model or a system of systems? How do we embed data resulting from the use phase into the DT? And finally, how can we guide the data integration process when it comes to updating the DT?

The main contribution of this review paper is to analyse, illustrate, and guide the effectiveness of different data integration approaches. The article explores the various data integration levels during the creation, use, and updating phases of UDTs, considering real-world implementations at the urban scale (namely city and district scale). To the best of our knowledge, this review paper is the first to offer a comprehensive analysis of data integration levels throughout the UDT lifecycle, comparing their advantages and limitations and illustrating them with current operational implementations. However, this review is subject to certain limitations, notably with regard to the selection of UDT projects to be investigated. For instance, we are unable to document all the initiatives since several initiatives are using the term digital twin, whereas in reality this is simply a digital model that fails to reflect the full UDT functionality and features. Thus, a series of criteria were applied to filter the different initiatives included in this study, in line with the levels of integration across the lifecycle.

First, the initiative's scale is the city, precinct, or district scale. Second, initiatives need to clearly emphasize the geospatial component and have geospatial data, notably the 3D city model, which is a key component in the creation of a UDT. In addition, the UDT project is operational and integrates a variety of data based on the use case requirements (i.e. air quality data, energy data, social data, to name a few). Finally, the UDT maturity model is also taking into consideration (e.g. based on the CITYSTEPS maturity model with an emphasis on stages from 3D static stage to real-time synchronization and autonomous implementation stage (Haraguchi, Funahashi, and Biljecki 2024)).

While UDT has many inherent technical challenges, our focus is on data integration within UDT. Our choice to address this major UDT technical challenge is that integration and interoperability are still considered the first issues in practice, especially that the complexity of data has increased under the umbrella of UDT as various multisource data and models are integrated. In this regard, the paper is structured as follows: Section 2 defines data integration and reviews related issues in 3D GIS (Geographic information system) and UDTs. Section 3 explains the levels of data integration. Section 4 investigates the challenges and potentialities of data integration levels. Section 5 represents and analyzes the levels of integration into the lifecycle of DTs. The same section exemplifies the different levels according to some ongoing implementations in practice. Section 6 highlights the main findings and gives a data integration overview. Section 7 concludes this work and gives a glimpse of future perspectives. These sections will enable the UDT developers and users to clearly understand the fundamental mechanisms and approaches of integration that rely on clear semantics and specific requirements and to prove the relevance of data integration levels based on different scenarios showcasing their applicability through some academic implementations.

2. Data integration and related issues

Many urban applications focus on data integration as an intrinsic phase in 3D city modelling (Stoter et al. 2020). Although the challenges related to multisource data integration have been recently tackled in different studies with a strong focus from the geospatial community, the challenges of data integration evolve increasingly under the auspice of the UDTs. Historically, spatial data infrastructure (SDI) has been developed at different levels and scales to assist effective data integration, usage, and sharing among different policies addressing technical and non-technical integration issues (Hu 2017; Janowicz et al. 2010; Kotsev et al. 2020). However, the actual SDIs (i.e. INSPIRE and the Open Geospatial Consortium 'OGC' catalogues) did not yet meet the maturity required for managing UDTs data integration challenges. Recent studies have suggested incorporating SDI within UDTs (Knezevic et al. 2022) proposed the implementation of an extended catalogue system in Germany, leveraging metadata and the open-source catalogue CKAN[1] software package. Another related work established an UDT and explored the development of an SDI-based energy domain using OGC standards to manage urban energy building data (Santhanavanich et al. 2022).

Furthermore (Noardo 2022), proposed a detailed and practical workflow for suitable multisource data integration, starting from defining the requirements based on the use case to the update and writing of metadata. The overall workflow can serve UDT implementations as it defines guidelines to properly understand and handle the fundamental integration issues based on the data coming from practice. For instance, the possibility to use data originally acquired for a specific use case in another application and vice versa is one of the benefits that UDTs have brought forward by focusing on integrating multiple domains.

Before we outline the main challenges of data integration, let us clarify the concept of data integration and specify what it covers. Indeed, data integration is acknowledged as the process of gathering and homogenizing multiple datasets from different sources depicting the same physical world feature and consolidating them into a unified and consistent format (Abdalla 2016). However, this definition is much more related to data fusion, merging, and harmonization through the encoding format and common database schemas (Li et al. 2020; Weil et al. 2023). These phases are regarded as components of data integration. Some relevant methods are available in the literature to merge and harmonize the data from various sources. Nevertheless, effective integration is partially achieved due to the lack of metadata, the differences in spatial and temporal scale, the geometry mismatch, and the lack of clear semantics, as many stakeholders are involved while moving towards the implementation of UDTs. Indeed, rich metadata should guide data integration, along with the definition of the use case and application requirements. Moreover, data integration is interlinked with the modelling paradigms, the storage, and the intended use. In this regard, data integration is conceived as the process of combining, transforming, and connecting multisource datasets into a single shared data model, database model, or one-stop platform, ensuring an accurate and insightful representation of the real world. It also takes into consideration, in the urban and geospatial scope, the enrichment process to create a complete and detailed 3DCM, enhancing further urban applications.

The integration within the UDT scope is particularly complex due to the lack of a common technical framework for data integration. On top of that, the integration focus is different when considering the lifecycle of DTs. Some authors focus on the data integration in the creation phase, which basically starts with the creation of 3DCMs where geospatial urban features need to be incorporated (DImitrov and Petrova-Antonova 2021; Fricke et al. 2023; Khawte et al. 2022). While other authors focus on the integration of specific data sets that could be dynamic, i.e. near real-time data from sensors and IoT data, or simulation data that, in turn, are integrated into the 3DCMs for further urban analysis (Chaturvedi 2021; Santhanavanich and Coors 2021).

Another data integration focus that gains interest while creating and updating UDTs is the integration of BIM and geoinformation, known as GeoBIM (Noardo et al. 2020).

When integrating various datasets, a considerable number of challenges need to be addressed, namely, syntactical, structural, and semantic levels of interoperability (IEC 2021)). The Open Geospatial Consortium (OGC) has formulated international standards and data models to tackle interoperability and, therefore, data integration issues (i.e. CityGML/CityJSON, SensorThings API (STA), OGC API features, to name a few). Thus, open standardized data models and exchange formats are required to enhance data interoperability and integration. Both (i.e. integration and interoperability) are technical barriers towards 3D city modelling and, more recently, UDTs (Billen et al. 2014; Döllner et al., n.d.; Kolbe et al., 2020; Kolbe and Donaubauer 2021). These challenges are interrelated. This means that data interoperability is a requirement as well as a part of data integration. Although data integration aims to foster usability and tackle the incompleteness of the datasets, data quality, availability, and conversion are serious challenges that hinder effective data integration. On the other hand, fundamental issues (such as semantic heterogeneities and information loss, limited or lack of metadata, geometry issues, standards inconsistencies, etc.) still require consideration while moving towards the implementation of UDTs. Further, understanding the semantic differences between each standard and each domain is vital to breaking down the data silos and reducing the complexity of existing standards (Kiourtis, Mavrogiorgou, and Kyriazis 2024).

In short, creating complex and connected systems such as UDTs requires integrating various datasets from different domains. Indeed, effective integration of various datasets is a key technical driver. However, the main question that should be answered is: what are the data integration levels, and what is the appropriate level of data integration based on the UDT lifecycle?

3. Conceptualization of data integration levels for UDTs

The heterogeneity of the technological framework and architecture to implement UDTs is as diverse as the data available in practice. However, while creating UDTs, cities reuse the existing data and generate new data sets through data integration and enrichment.

The investigation of DTs implementations from practice deployed different levels of data integration based on the use case requirements (UDTs-based use case). Furthermore, there is no universal definition of what UDTs represent and few related works define the

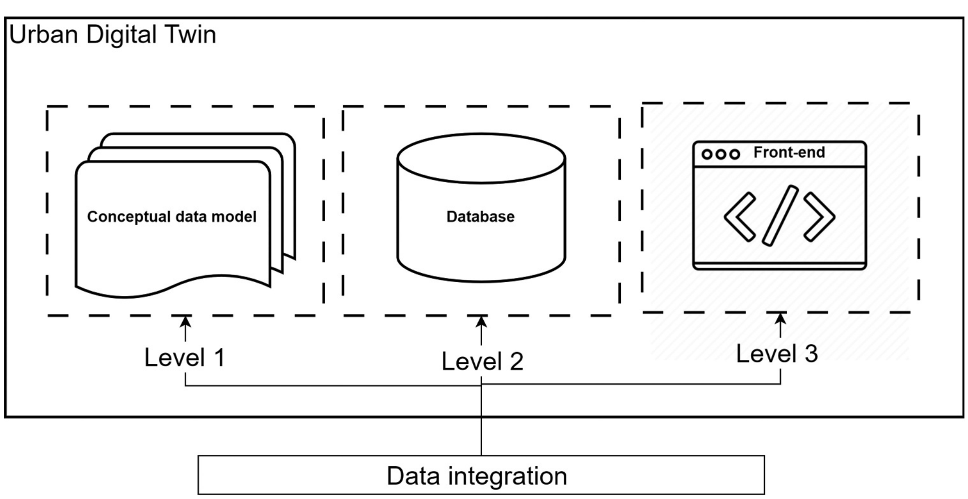

Figure 1. Levels of data integration: possible ways for integrating data into UDT.

minimum technical requirements to develop a true 'UDT'. Thus, understanding the data integration approaches is fundamental since these levels in their turn pose their own challenges.

Based on the literature and the classification of a significant number of UDT implementations, three levels of data integration were defined: the conceptual data model (CDM), the database level (DB), and the front-end level (see Figure 1).

3.1. Conceptual data model level (CDM)

The conceptual data model (CDM) is defined as the schema that structures the data and the relations between entities and attributes.

In contrast, from a technical perspective, existing data models and standards are deployed; either they are extended to cover new themes and new features, or multiple data models or subsets of data models are merged based on a new unified data model based on the DT requirements. Thus, performing data integration at the conceptual data model is the upper level of integration.

At the time of writing, there is no official conceptual data model designed to cover all UDT requirements. This offers the user the possibility to alter and extend the existing conceptual core model to integrate objects that are not already represented by any other class. Indeed, integrating at this level implies adapting the core model using extensions provided by the current CityGML standard and its encodings (JSON and 3DCityDB). In fact, CityGML[2] is the most used data model for semantic 3DCMs. Its conceptual data model has been considered a data hub and integrative platform for urban and geospatial data for decades, allowing data integration and the creation of a comprehensive Urban Information Model (UIM). They are therefore considered the best candidates for introducing the 1st level of data integration in the context of UDTs serving as an abstraction of the physical world and an integration basis. This data model is generally extended to cover and incorporate different data for generic purposes (i.e. adapting the conceptual data model to a national or regional context) or to support specific use cases and applications through ADEs. This level offers a direct feed into the database.

3.2. Database level

The database level could be defined as the process of converting the data to be integrated to match the conceptual data model and the database schema. In fact, the CDM is not edited; otherwise, we will be talking about the 1st level of integration. The database level allows for feeding or updating specific classes or attributes without altering the semantic consistency of the data model. At this level, the data is adjusted for database feeding. ETL (Extract, Transform, Load) processes are deployed to transform the data to meet the data model structure and schema (Sreemathy et al. 2021). For this purpose, the model is unchangeable, but rather the data needs to be adjusted in line with the data model.

3.3. Front-end level (client side)

This level represents the UDT's front-end. Integration at this level could be performed via various interfaces and tools (i.e. GIS tools, web applications, geoportals, dashboards, linked data and game engine applications, to name a few) (refer to Figure 1).

Web-based applications show their capabilities in visualizing and handling complex 3D data on modern browsers. While moving towards UDTs, 3D data integration on the Web has been flourishing. This data integration level represents the most common approach to multisource data integration in the current UDT implementation, as previously investigated in Jeddoub et al. (2023). It allows the data to be integrated into common interfaces and handle different data exchange formats, achieving syntactic interoperability. Furthermore, many cities implement a distributed architecture for data integration and management and disseminate the 3D content 'as data-based services' where many stakeholders can interact while maintaining their own data. Hence, the data can be served and visualized on the web by various parties without following a data integration process at the data model level. In addition, the use of web-based platforms provides API support for developers to optimize data access more effectively with the availability of current web-friendly standards releases.

The ambition towards fostering open data through the web and making it accessible to multiple stakeholders enhances the need to investigate the development of web UDT platforms. However, combining various data and models into a one-stop platform may lead to some inconsistencies. For example, data coming from practice has diverse heterogeneities ranging from the semantic representation to the syntax, the structure, and the geometry. Furthermore, the data may encompass various spatial and temporal scales. However, from a technical perspective, current platforms based on WebGL JavaScript libraries, such as CesiumJS,[3] increasingly support multi-format data integration. Indeed, Cesium, the common open-source platform, is used to integrate and visualize the required data. The platform supports several formats, and the data basically goes through a data conversion process. Nevertheless, integration might be affected at the semantic level. Indeed, both semantic and syntactic interoperability are extremely valuable while developing collaborative platforms such as UDTs. In parallel, the semantic web approaches (i.e. linked data) highlight their strengths in performing data integration in the built environment.

The data integration at the front-end level based on the web still represents some challenges in practice that are addressed within the geospatial community, namely through data standardization and interoperability. While web-based platforms have shown their capabilities in handling various 2D and 3D geospatial content, game engines, in turn, have attracted considerable attention as effective tools for UDTs. Although browser-based platforms such as CesiumJS and ArcGIS Maps SDK for JS[4] facilitate user accessibility as well as enhancing comprehensive support of geospatial data, their visualization and interactive capabilities are still limited, especially when operating on large urban scales such as cities. This is where game engines have found room for development.

Game engines like Unreal Engine (UE) and Unity bring new opportunities to access and visualize real-world data represented through urban and geospatial data with high visual rendering in an immersive way. Furthermore, the game engine plugins available in practice, such as the UE plugins for Cesium Web Globe and Esri's Web Globe, consolidate the link between 3D geospatial technologies and game engine platforms. Nowadays, game engines are considered a trend for data integration, particularly under the UDTs umbrella. From data standardization and interoperability perspectives, an early exploratory study was conducted to test the usability, compatibility, and viability of existing geospatial OGC standards, namely, 3D Tiles, I3S, and 3D GeoVolumes API, with game engines (i.e. UE and Unity) based on Cesium and ESRI's plugins (Würstle et al. 2022). The authors illustrate through three use cases the usability and applicability of the OGC standards in the framework of game engines. Furthermore, challenges faced in the implementation of the use-case-based OGC standards were documented. These limitations need to be studied in a comprehensive way since they affect the final model.

In conclusion, various architectures are proposed in practice and are mainly based on a web browser platform, namely Cesium and ESRI solutions such as ArcGIS online and JavaScript for ArcGIS. Cesium is widely used in academic UDT initiatives compared to ESRI tools. This is certainly due to their open-source initiative. The browser-based geospatial platforms show their capabilities in accessing, integrating, and visualizing 3D content. Although they have great support for geospatial OGC standards, they are still limited in terms of effective rendering and interactivity. Hence, there is a need to further explore the potentialities of game engines.

4. Data integration levels: pros and cons

The data integration levels have their pros and cons when developing an UDT. This section discusses the strengths and limitations of each data integration level and proposes guidelines for users in terms of their effectiveness.

Table 1 highlights some of the advantages and disadvantages of the data integration levels. The following list of pros and cons is not exhaustive, and further improvements and drawbacks may emerge in the future.

According to Table 1, level 1, which performs integration at the data model, provides a range of advantages, including semantic interoperability and compatibility, taking advantage of the model's extensibility and scalability, and the possibility of schema validation after

Table 1. The pros and cons of data integration levels.

	Level 1	Level 2	Level 3
Pros	+Semantic interoperability + Model extensibility + Coherent and harmonized data model + Schema validation + Data import "as is" (no conversion needed) + Direct load to database	+ Straightforward data model compatibility + Native support of city objects (i.e. CityGML urban objects) + Use of generics attributes (i.e. CityGML generics) + New versions of data model enhance direct data integration.	+ Syntactic interoperability + Separated and independent database management systems (asynchronous update cycle) + On the fly simulations (based web services and processes) + No data conversion + Various data format support
Cons	- Model complexity - Compliance with the Application Domain formalities (i.e. CityGML ADEs) - Large data model (huge number of classes, attributes, properties) - Compatibility with new data model versions. - Software support for extensions is still limited.	Preprocessing needed - Data loss (semantic loss) - Mapping between entities and attributes could be tricky.	- Heavier bandwidth. - Limited interactive client-side

extending the data model. Furthermore, data can be imported directly without the need for a data conversion process. In addition, data is directly fed into the database, or more specifically, when the DT's technological framework relies on databases for data management and storage.

While there are significant advantages offered by this level of integration, there are also a series of challenges that should still be addressed. Extending a data model certainly allows new classes, entities, and attributes to be managed, but this may increase the complexity of the data model (for example, adapting a CityGML model presents a range of challenges given its complexity). Besides, extending a data model is not done in a random way, but rather a certain logic and formalism must be followed and maintained. Extending a model in general creates heavy files, and unless these files are handled through a database, it would be rather cumbersome to use them for advanced analysis. Versions of the model may change over time, and new classes may be added, updated or deleted. Changes come to support direct integration (level 2) with data related to specific applications or generic contexts (example of the 'Dynamizers' concept of CityGML). Extended models need to be supported by a wide variety of tools. Unfortunately, this is still limited. Hence, integrating at the data model level is not simple, and solid knowledge is required to achieve the intended goal.

Considering the 2nd level of integration, the related advantage is the straightforward data model compatibility. The data model supports, by default, the data to be integrated. For instance, the current implementations of UDT that are based on the CityGML data model have native support for almost all urban objects. Integration could also be performed using generic objects and attributes provided by the data model. Moreover, the new version of the data model enhances direct integration of the data into the model without further model modification or extensions. However, this integration at the database level might require data preparation and data conversion. Indeed, the data may not be used as delivered, but some conversion workflows are mandatory to meet the model requirement. This leads to semantic data loss, and more specifically, during the mapping process between entities and attributes, which could be tricky.

We can state that levels 1 and 2 are both traditional levels of integration that have existed for decades under the auspices of many similar concepts. However, level 3 has been gaining high interest with the emergence of the UDT. Level 3 supports data integration through software and platforms. The integration at this level enhances syntactic interoperability. These architectures are based on separate and independent database management systems, which facilitate the DT update. In addition, recent implementations show the possibility of running on-the-fly simulations based on web services and processes. The integration does not necessarily require a data conversion process since the current platform allows the handling and integration of various data types and formats. Level 3 also has some limitations, namely limited interactions, slow server response times, and poor rendering performance. The DT implementations that integrate the data at level 3 are generally tested in small areas (i.e. precincts and district levels). However, incorporating data using the same workflows at the city level requires optimizing the management system and visualization tools of such massive data. Recently, many studies have been carried out to effectively render and represent 3DCMs on the web. Hence, OGC standards and formats, such as 3D tiles and Indexed 3D Scene Layer (I3S), were developed to facilitate the streaming of these 3D models on the web. However, this

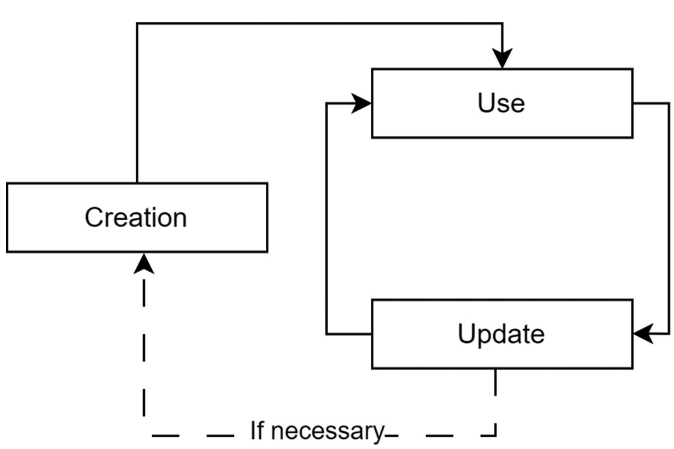

Figure 2. The urban digital twin lifecycle stages.

format conversion from, for example, the CityGML 3DCM as a common input data to UDT to other web-friendly formats led to semantic data loss (Padsala et al. 2023). Furthermore, the conversion from CityGML to the web streaming format raises update issues, i.e. the integration of datasets coming, for example, from simulation software needs to be processed whenever new results are made available, and the updated CityGML model must be again converted to the required web data exchange format for visualization.

5. Unpacking and illustrating data integration levels according to the DT lifecycle

In the framework of this review paper, we intend to explain the data integration levels based on the DT lifecycle stages and analyse them accordingly (refer to Figure 2). The lifecycle is defined by three primary stages: creation, which involves the generation of a UDT; usage, which focuses on the deployment of the fundamental component of the UDT namely the 3DCM in simulation applications and urban analytics workflows; and lastly, the update phase, which serves to continually maintain the UDT. We therefore conduct a more detailed analysis of these data integration levels according to the lifecycle stages, data types, and current and possible scenarios. We have included the scenario dimension to highlight the various integration-level options currently available and those that might be feasible in the future. This scenario exemplifies an UDT implementation which uses multiple data sets, and which can incorporate those data sets based on different levels of integration. By default, we have covered all the possible combination of these approaches, leading to the identification of different scenarios. These scenarios will be further illustrated whenever possible by current initiatives and implementations.

5.1. Creation of UDT

The creation phase is defined as the first step towards the implementation of UDTs. This phase consists of the conceptualization of all relevant components, considering the core dataset to be integrated (usually geo-data), the modelling approaches, the required level of detail, and the representation and structuring of both data and models.

Given that, in practice, the 3D city model is a part of UDT and not UDT per se, we will focus on this creation phase on the enrichment of 3DCMs as an integration foundation for UDT. In practice, creating UDTs is often seen as building a semantic city model or connecting different data models and systems into one unified model. The creation phase in this study implies the enrichment of existing city models or systems to meet the DT's requirements since most of the implementations are not built from scratch. This is achieved concretely through data integration.

To create an UDT, data integration can be performed at different levels, as defined earlier in Section 3. Our focus is not on data-to-data integration (also known as data fusion), but we are particularly interested in data-to-model and model-to-model integration. In addition, the seamless integration of various urban city objects to create a complete and comprehensive 3DCM needs to be considered as well.

To clearly understand the three levels of integration, we extract the common scenarios for creating an UDT based on the literature. Figure 3 gives the various current scenarios of data integration levels in the creation phase.

Scenario A: This involves integration at level 1. This requires changing the model if it does not handle the integrated data by default. In this regard, CityGML is widely adopted as a data model for semantic 3DCMs,

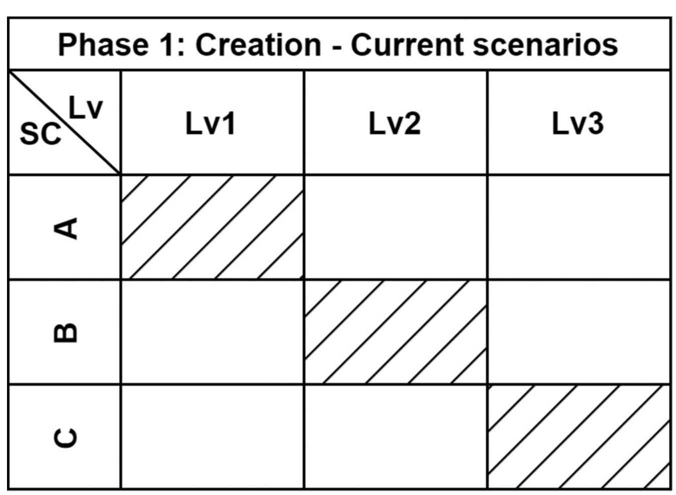

Figure 3. Current scenarios of data integration levels in the creation phase.

and recently, UDTs. Thanks to its common urban object native support, its extensibility and its interoperability, it is considered a basis for data integration to create a comprehensive and accurate UDT. This creation could be achieved by augmenting its data model with new and additional concepts using generic ADEs. For instance, generic ADEs are implemented to adapt the data model to support a national standard or to cope with specific needs.

Most CityGML models are based on version 2.0, which probably uses ADE to handle and integrate new datasets and thus tackle new problems. With the new version 3.0, this integration has changed, for example, the 'Dynamizer concept' has a direct compatibility with the core model, thus enabling data integration at level 2 according to our definition. In this regard, new versions of data models may introduce new classes and attributes to the schema or remove some parts of them to fit the UDT requirements. This ultimately affects the level of integration.

Scenario B: Integration at the database level involves the use of an ETL process to feed data into the model. The database level is commonly used in UDT implementations based on the native data model or through the support of generic objects and attributes.

Scenario C: Data integration at level 3 is relevant in the creation phase. The integration takes place at the front-end level. The model remains untouched, and the databases are completely independent. The data is integrated using a web architecture or game engines.

Table 2 further illustrates the different levels of data integration. It provides current illustrative DT implementations that are in the creation phase and use different data integration levels.

Many examples might be used to illustrate these levels during the integration phase. To illustrate our statements, we discuss below several related research under the prism of our classification.

Based on Table 2, Level 1 is used in the preliminary phase of creating the DT of the City of Vienna (Lehner et al. 2024). To support the long-term vision of creating a DT of the city, an extension of the CityGML 2.0 data model profile is being performed to meet the current and future requirements of the municipality. The data model is augmented with additional features using ADE to create a complete 3D model of the city. For level 2, a study case was performed within an Australian DT pilot project (Diakite et al. 2022). The authors integrate the provided spatial data sets using a database approach. They process, incorporate, and store existing 3D data in the 3DCityDB using PostgreSQL and PostGIS database, the main core of their DT architecture. The data went through a mapping process to meet the CityGML classes and attributes since the data were provided in different structures. Hence, the direct import using the official 3DCityDB Importer tool of the data sets to the database was tricky. In this regard, a series of Python algorithms were developed to reach the integration goals. To illustrate the level 3, recent related work designs a socio-technical conceptual framework that illustrates the need to incorporate human sensing data (i.e. participatory data) into UDTs (Lei, Su, and Biljecki 2024). In their case study, the authors implemented their UDT prototype based on multiple open data sources (static and dynamic data), namely, crowdsourced data (e.g. Open Street Map data), participatory data (social sensing data), and environmental data. Data integration was performed using Cesium. The platform is based on a browser-server architecture, fostering data interoperability and system compatibility with current and upcoming input data and systems. The data went through a data conversion process based on an FME Flow. The platform supports several formats, such as 3D tiles and JSON-based

Table 2. Examples of data integration levels (lv) in the creation phase using different data inputs.

	Creation phase		
Data input	Lv1 (Scenario A)	Lv2 (Scenario B)	Lv3 (Scenario C)
Static geospatial data	- CityGML 2.0 ADE of the Vienna City (Lehner et al. 2024)	- Mapping the data and their attributes into the 3DCityDB tables (Diakite et al. 2022)	- Cesium Ion (Lei, Su, and Biljecki 2024) - Cesium JS, VR and JS libraries (La Guardia and Koeva 2023; La Guardia et al. 2022) - COVISE (Collaborative Visualization and Simulation Environment) (Dembski et al. 2020)
Other static data	- IFC-CityGML data model (Li et al. 2019, 2020)	- CityGML 3.0 (space concept) (Kutzner, Chaturvedi, and Kolbe 2020)	- Integration of both BIM data and geodata in the same front-end (i.e. Esri ArcGIS CityEngine,[5] ArcGIS GeoBIM,[6] virtualcitySYSTEMS[7])
Dynamic data	- Dynamizer ADE (CityGML 2.0) (Chaturvedi 2021; Chaturvedi et al. 2017) - Extending 3DCityDB by Dynamizer ADE (CityGML 2.0) (Chaturvedi, Yao, and Kolbe 2019) - Dynamizer (CityJSON 1.1) (Boumhidi, Nys, and Hajji 2024)	- Generic attributes (Chaturvedi 2021) - Dynamizer module (CityGML 3.0) (Kutzner, Chaturvedi, and Kolbe 2020)	- CityThings (connection between the SensorThings API (STA) and the 3DCM using gml-id) (Santhanavanich and Coors 2021) -COVISE (Collaborative Visualization and Simulation Environment) (Dembski et al. 2020)

formats, as well as Cesium Markup Language (CZML), that represent human sensing data interlinked with the geospatial component. Another related work proposes a 3D data fusion workflow based on Virtual Reality and JavaScript library-based 3D WebGIS solutions (La Guardia and Koeva 2023; La Guardia et al. 2022). In their work, the authors handle and integrate different and complex geospatial input data, namely point cloud data, 3D models, and BIM models.

The same logic applies to BIM-GIS integration, which is considered one of the fundamental data integration challenges while moving towards implementing such a transdisciplinary concept as UDT. This integration could be performed at level 1. For instance, a study focuses on the precinct scale and develops a semantic precinct information model (PIM) using multisource and various data (Li et al. 2019, 2020). Based on the conceptual and unified data model, the authors implemented a relational spatial database to manage the building model. The integration at level 2 is feasible by mapping IFC classes to CityGML classes and attributes, as well as with the last CityGML 3.0 (Kutzner, Chaturvedi, and Kolbe 2020). Data integration is achieved at level 3 via GIS and urban tools (i.e. ESRI CityEngine, ArcGIS GeoBIM, virtualcitySYSTEMS).

Integrating dynamic data (such as sensor data) highlights the key features of UDT. Indeed, this data integration within UDT can be achieved by extending the data model concept, as seen in 'Dynamizer ADE' for version 2.0 (Chaturvedi et al. 2017). Alternatively, integration can occur at level 2, exemplified by the use of the thematic module 'Dynamizer' in CityGML 3.0 (Kutzner, Chaturvedi, and Kolbe 2020) or sometimes using generic attributes (Chaturvedi 2021). At level 3, integration happens through system architectures, notably by ensuring a connection between the server handling the sensor data and the 3DCMs (Santhanavanich and Coors 2021).

Herrenberg's city-scale DT prototype is a project that illustrates the heterogeneous data integration (i.e. static and dynamic data) in the creation phase, where various existing volunteered geographic information and a 3D city model are used. Furthermore, thematic data such as mobility data from mobile applications and GPS (i.e. movement traces of bicycles and pedestrians) are used. The integration is performed at the front-end level. The project uses COVISE (Collaborative Visualization and Simulation Environment), an extendable distributed software that supports multisource data and sensor network data as well, resulting from a computational model for air pollution simulation.

In this regard, data integration in the creation phase could be performed according to different levels of data integration. As discussed earlier, there are current scenarios that are implemented in practice. However, expected theoretical scenarios in the context of creating UDT need further investigation (see Figure 4). Indeed, it is possible to combine different levels of integration. This leads to the following integration scenarios.

Scenario D: Integration in this case might explore the 1st and 2nd levels of data integration. For instance, road data is provided and needs to be incorporated into the UDT. Level 2 might be used to convert the data to meet the requirements of the data model, which could also be

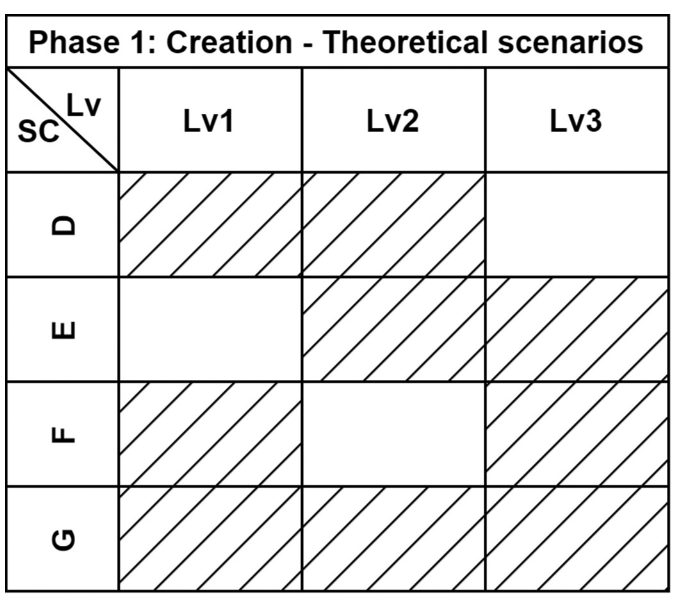

Figure 4. Theoretical scenarios of data integration levels in the creation phase.

extended and modified (level 1) to receive the transformed data.

Scenario E: Levels 2 and 3 are used in this context. For instance, the data model will natively handle the data to be integrated, but additional data is not supported. Thus, integrate them using web architectures or game engine workflows.

Scenario F: It involves the use of levels 1 and 3. Data could be integrated based on the 1st level, and further data in the same UDT implementation uses the front-end level to integrate the data.

Scenario G: This scenario is the most faithful to the definition of a generic DT, in which a DT has a concrete data model, a database that manage the corresponding objects, and, finally, one or several client applications. With this in mind, the DT is neither a file- nor database-based system but rather can be assimilated into a three-tier architecture (model/DB-server-client).

Implementations D, E, F, and G are proposed in this work as a combination of different levels to integrate data into UDT. Yet, such implementations are hard to find and should be investigated in the future.

From practice, the creation phase is mainly focused on the data enrichment of the available 3D models to generate an accurate and complete digital representation. The enrichment is performed based on the different levels of integration. Based on our investigations in the creation phase, we can clearly conclude that level 2 based on the CityGML 2.0 data model is the common data integration level. Level 1 is explored to meet the national DT requirements using ADE in this preliminary phase. However, this needs further investigation in the future. Level 3 is explored as well by connecting different input data at the front-end level.

5.2. Use of UDT

The use of the UDT corresponds to its operation in a specific use case to address a particular application. In practice, this stage can sometimes be confused with the creation phase, particularly when the creation of a DT is based on a use case, also called a UDT-based use case. Nevertheless, the use of UDT is defined as the process of performing simulations and urban analysis. This phase is closely related to the use case requirements, where input data is required to generate new data sets or models, which in turn will be re-injected into UDT to support decision-making. The produced data sets in this phase might be incorporated into the DT following various data integration levels. While these integration levels have been experienced with 3DCMs for multiple use cases and applications, this section will tackle this integration aspect within the context of UDT available implementations.

Within the geospatial community, 3DCMs form the foundation for these data integration levels in the UDT scope. In this regard, and considering current and ongoing UDT implementations, we intend to outline and discuss the levels of integration identified in the literature.

Scenario A: This scenario is also known as the traditional way of integrating data. It involves the extension of the CMD to manage and store the data generated from the simulation tools into the 3DCMs. Indeed, ADEs, namely Energy ADE and others, are used to perform this integration task (Agugiaro et al. 2018). Moreover, with the growing interest in breaking down data silos, defining a standardized open data model is thus of great interest. For instance, studies investigate the possibility of designing a Food-Water Energy ADE (FWE ADE) (Padsala et al. 2021). The approach is based on extending the CityGML data model to support the integration of food, water, and energy data and simulation workflows. The approach is tested at different spatial levels with different input data. The results are stored in the added module as defined in the FWE CDM. While moving towards sustainable cities, attempts to manage and store multidomain data are still limited to a few implementations.

Scenario B: Level 2 is explored in this scenario. The model handles by default the simulation outputs, measured data, or any other data produced by simulations and urban analysis or provided by third parties.

Scenario C: This level requires a system architecture that allows bottom-up data integration of urban information and 3DCMs (associating separate database management systems) as well as their visualization.

Scenario D: This approach involves data integration based on levels 1 and 2. The combination of levels 1 and 2 consists of adapting the data model to integrate the data and, at the same time, using level 2 to enrich the model where applicable. For example, extending the CityGML model to incorporate simulation results and using generic tables to integrate further data.

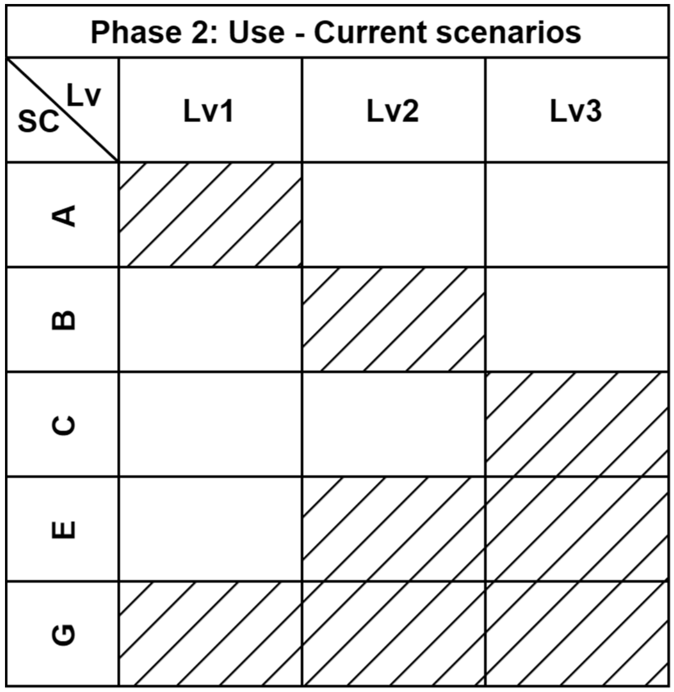

Figure 5. Current scenarios of data integration levels in the use phase based on an energy use case.

Table 3. Examples of data integration levels in the use phase based on an energy use case.

	Use phase		
Data input	Lv1 (Scenario A)	Lv2 (Scenario B)	Lv3 (Scenario C)
Energy simulations	- Energy ADE (Würstle et al. 2020)	- Extended 3DCityDB with Energy ADE tables (Rossknecht and Airaksinen 2020)	
Measured data	- "Dynamizer" ADE (Chaturvedi 2021)	- Extended 3D city DB with generic tables (Würstle et al. 2020)	- CityThings concept (Santhanavanich and Coors 2021)
	Lv2 and Lv 3 (Scenario E)		
Static properties and time series datasets simulation	-Static propreties/Attributes (generic attributes) -CityThings concept (Santhanavanich et al. 2022)		
	Lv 1, Lv 2 and Lv 3 (Scenario G)		
2D geometry and attribute data/Temporal sensor data/Urban building energy data	- Energy ADE schema (Building Energy and Climate Atlas)- Enrichment of CityGML data model with attributes of building function and year of construction- OGC API – Features- STA service - OGC API Processes services (Santhanavanich et al. 2023)		

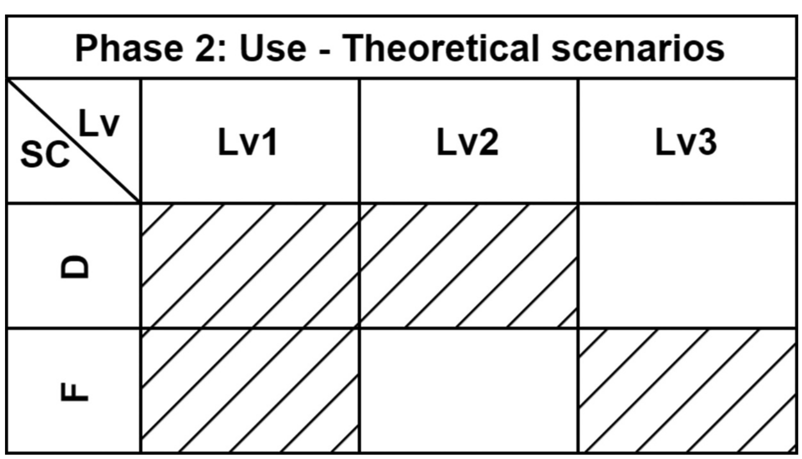

Figure 6. Theoretical scenarios of data integration levels in the use phase.

Scenario E: Levels 2 and 3 are used to integrate data into the UDT. This scenario takes advantage of level 2 since the data model might by default support data integration for a given use case; at the same time, it uses level 3 to enrich the UDT with various urban data.

Scenario F: This scenario involves the use of both levels 1 and 3. This implies that a data model is adapted for a specific use case and an additional system architecture is designed to support the integration.

Scenario G: This scenario implies the use of the three levels of data integration. Each level is deployed to integrate a specific data type on the same UDT platform.

To illustrate the current scenarios, we adopt the energy use case. The use case was selected on the basis that many research studies have already identified and tested the different levels of integration. Thus, we extracted the current scenarios (refer to Figure 5) and analysed the references linked to this UDT based on data types (refer to Table 3) and identified theoretical scenarios (see Figure 6).

Table 3 identifies the different levels of data integration. It provides illustrative DT implementations that are in the use phase. The implementations use different data integration levels.

Based on Table 3, current scenarios are identified as follows:

Scenario A: Based on CityGML 2.0, an energy-related urban data model is defined to calculate and enrich the data model with building energy information required for energy simulations (Würstle et al. 2020). Hence, Energy ADE has helped in addressing data interoperability issues. In addition, 'Dynamizer' ADE is also used to manage measured data in the energy use case (Chaturvedi 2021).

Scenario B: Extending the 3DCityDB to support Energy ADE allows us to perform data integration at level 2 (Rossknecht and Airaksinen 2020). For instance, energy information is stored in 3DCityDB, where Energy ADE tables are already embedded. The aim of this database approach is to directly feed and update the database tables from the simulation results without the need to alter the CMD. The results are written back to the database, escaping the challenges related to the file-based approach. Alternatively, integration at Level 2 might be carried out through generic tables (Würstle et al. 2020).

Scenario C: For instance, the 'CityThings' concept enables managing and integrating dynamic sensor data and 3DCMs provided by different parties in a separate management system (Santhanavanich and Coors 2021). The concept connects sensor data from SensorThings server and the 3DCM. The approach addresses the challenges related to the direct storage of high-frequency measured data into the CityGML data model (either for file- or DB-based approaches). In fact, level 3 is designed when the data model does not allow integration or represents some limitations; hence, the adoption of system architectures and the APIs standards at the front-end level facilitates integration and maintenance within the DTs framework.

Scenario E: An illustrative use case of this level of integration is provided in the work done by Santhanavanich et al. (2022). The authors propose a method to tackle the traditional data integration workflow of the energy simulation results into the 3DCM before converting the new CityGML model to a ready-to-web stream format. The aim is to optimize the data management of 3DCMs using SDI through web applications. In their SDI approach, the authors use a range of OGC web services standards, namely the SensorThings API, the OGC API 3D GeoVolumes, the OCG API Features, and the CityGML standard. The simulation outputs are handled differently; the static building properties are integrated directly into the CityGML data model (which refers to level 2). However, the spatiotemporal data is stored and managed in an external database. This was performed at level 3 based on the STA exchange format specification by ensuring a connection link to the OGC web services.

Scenario G: This scenario implies the use of the three levels of data integration. Each level is deployed to integrate a specific data type on the same UDT platform. This scenario was recently implemented in a study by Santhanavanich et al. (2023). The authors illustrate a web-based API workflow to integrate and handle geospatial data, 3DCMs, building energy data, and simulation results based on a web service and OGC API standards. The aim of the approach is to create a single and harmonized data model combining geospatial and energy data that are in practice provided and maintained by different providers. The approach involves various data integration levels. Energy ADE is explored, which illustrates the 1st level of integration. Furthermore, the enrichment of the CityGML model is performed at level 2 through mapping approaches. Finally, the simulation results are integrated based on a modular web architecture based on OGC API standards, namely OGC processes and the STA. The overall web-based API architecture is tested during OGC Testbed 18[8] for two study use cases (i.e. Montreal and Helsinki). The aim of the testbed is to extend the geospatial spatial data infrastructure to support building energy data interoperability as well as setting the fundamentals of energy spatial data infrastructure. The SDI addresses interoperability issues and integration challenges through standardized web services and interfaces. Data visualization is also performed successfully using innovative tools to facilitate data access and analysis (WebGL and AR applications).

The testbed-18 pilot project aims to generalize the workflow to the entire city using a high-level developed prototype. Furthermore, the participants highlighted the potential of moving towards a generalized data model. This data model seeks to structure and consolidate both geospatial and energy datasets based on a standardized modular web services-based architecture.

Scenarios D and F are identified as theoretical scenarios. They have been less explored in the framework of integrating data into UDTs (see Figure 6).

Scenario D: Given that we are interested in the UDT-based energy use case, we were unable to find an illustrative study case that allows us to further explain the integration process.

Scenario F: In this case, a suitable integration-based energy use case was hard to find.

Based on the previous analysis, the use phase is closely driven by the use-case requirements. This phase encourages collaboration and convergence between various domains. However, the heterogeneity and complexity of the data in this phase raise data integration challenges. In this regard, data integration levels have shifted increasingly from file-based approaches to standard web-based API approaches. The new trend towards data integration at level 3 has been involved in recent UDT implementations. It is considered the most suitable level to integrate any type of data. However, the data, namely geospatial data, needs to be FAIR (findable, accessible, interoperable, and reusable), and domain-specific data should be provided in a standard-compliant format. In addition, the data needs to be well documented to facilitate the integration and maintenance processes with formatted metadata and sufficient technical documentation. This mapping of integration levels, as detailed above, can be extended to cover any use case. This analysis of the levels based on the use case enables us to identify the patterns of data or model integration according to the urban application. It is also worth noting that the use phase can serve as a guide for the creation phase. For example, in the creation phase, a semantic 3DCM was built based on level 1. In the use phase, it appeared that model enrichment was required to run the simulations. This will probably prompt level 1 or 2 of data integration to generate the necessary data for the use phase.

5.3. Update of UDT

The update phase is an extremely valuable step in the lifecycle of UDTs. This phase is unfortunately overlooked in current implementations. Nevertheless, it is always highlighted as one of the key challenges to be considered when maintaining DTs. The UDT should be conceived as a sustainable process and not as the final product. For this reason, the definition of integration

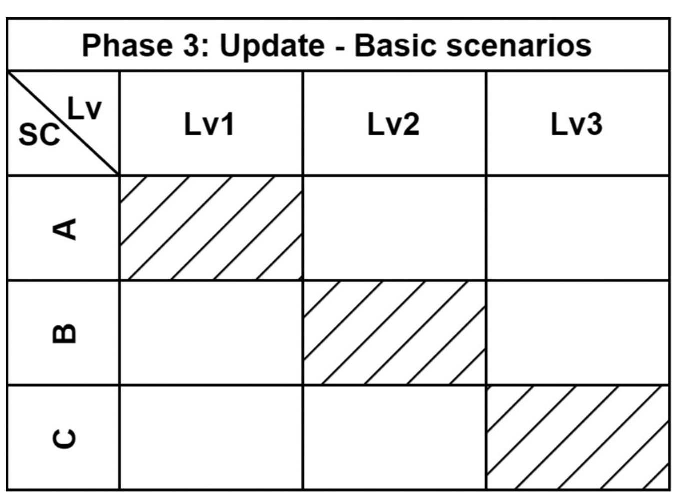

Figure 7. Scenarios of data integration levels in the update phase.

approaches is an essential requirement in the updating phase. That is how these different levels have their merits and demerits. Proper strategies and mechanisms to support the update of UDTs are still limited. This issue has already been experienced with 3DCMs. Although various DT parties (i.e. geodata providers, researchers, and municipalities) are aware of the necessity of data maintenance and governance. Current implementation still lacks concrete investigations. Given that performing updates is tightly related to data integration, we will discuss in this section the data integration levels in this phase of the lifecycle.

Updating an UDT according to the three basic levels as previously defined can be envisaged (see Figure 7); however, a scenario-based analysis, as conducted in the creation and use phases, is not currently going to be carried out, given that the update mechanisms are limited to a few DT prototypes.

One of the main characteristics of UDTs is that the virtual system is updated regularly through the bi-directional link between the real and digital worlds. Hence, data integration methods should be set to guide the user while integrating new data or information. The new data could be obtained from different sources and cover different themes. However, the appropriate level of data integration is guided by the previous phases (i.e. creation and use).

Data maintenance logic should consider the data, which is the fuel of the DT, and the DT components, which structure these data. In this respect, it is vital to distinguish between the updating of data, which may be handled by the data provider, and the updating of DT's components given a new set of data, which does not solely rely on the data but rather encompasses the techniques and methods applied in the development of these digital twins.

To illustrate, let us consider the updating of a 3D city model, which is one of the main components of Digital Twins. Possible updates in the creation phase are, for example, a new LiDAR acquisition, a new building created, a road removed, or new vegetation, to name a few. These data are generally updated by their providers (notably the cadastral department). Although this is done systematically, however, how do we update the UDT? There are several possibilities for a new LiDAR acquisition: (1) A new LiDAR acquisition implies a new model and, therefore, a new integration of the model in the UDT (re-modelling). (2) compare the new point cloud acquisition with the old one using change detection methods, then regenerate the 3D model. (3) compare existing models with the point cloud, then update the model. (4) compare based on model-to-model approaches and explore versioning of 3D city models. In all cases, these possibilities will require the reintegration of the 3D city model into the UDT. Furthermore, the update mechanism is related to the data integration levels used in the creation phase. Indeed, if the creation phase is based on the conceptual data model level, the update would be performed at level 2, given that the core model used in the creation phase handles these urban objects (see Figure 8a).

The same logic is applied to the usage phase. For instance, to update specific data that was integrated into the use phase (i.e. when new simulation workflows are run based on new datasets or parameters), the new outputs should be injected into the UDT, and the attributes need to be updated accordingly (refer to Figure 8b). Generally, level 3 is the most convenient data integration approach for dynamic databases, such as the one explored in the UDT in the use phase. This level seems appropriate for dynamic data such as the one measured from sensor data as well as the simulated outputs. The particularity of these data is that they are

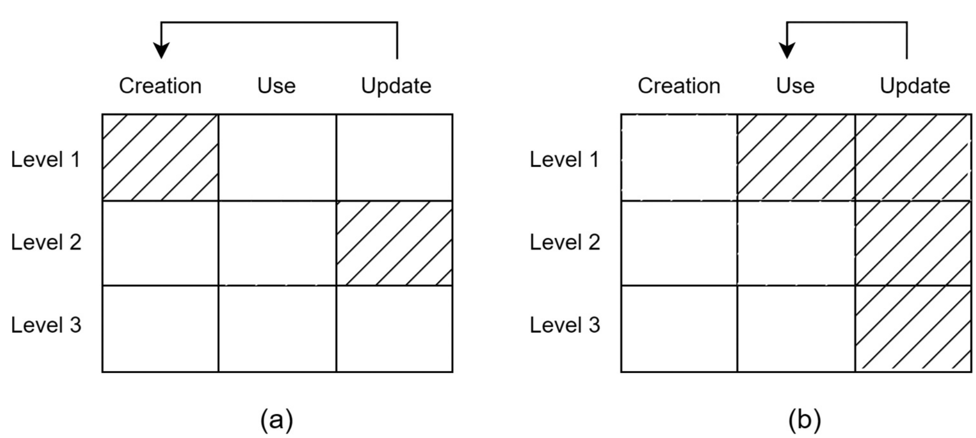

Figure 8. Data integration levels for DT updates: (a) at the creation phase, (b) at the use phase.

generally managed in separate systems in the current implementation. Thus, updating them is performed in a straightforward manner. Furthermore, level 3 appears to be a promising approach to data update since the data are maintained by their owners and shared as an asynchronous and up-to-date service based on a web-based architecture. This increases access, querying, sharing, and updating of the data. Furthermore, the update could be performed on the front-end side by means of user interfaces.

Although we have already outlined the basic update scenarios, in this phase, the main question to answer is which part of UDTs needs to be updated, so the data integration levels are decided accordingly. The data deployed in the update phase needs to conform to the UDT requirements. Furthermore, standardization is mandatory to detect changes in the data and facilitate their maintenance. It is also important to highlight that updates are usually more focused on the use phase than on the creation stage. This is because the whole purpose of a UDT is to create it only once, to use it in multiple use cases, and to perform updates when appropriate, especially as this is a costly process in term of human and technical resources and will depend on the availability of data.

6. Data integration levels: discussion and findings

Data integration is considered both a fundamental requirement of UDT and a key challenge that hinders its full implementation in practice. Different models, architectures and platforms are developed and named UDTs. However, they represent discrepancies in terms of how the data is integrated. To properly understand the data integration levels, we represent these levels according to the DT lifecycle inspired by the geospatial data lifecycle. Nevertheless, geospatial data are not the only kind of data to be incorporated into DTs; various domain-specific data can also be integrated (i.e. energy data, noise and air quality data, computational fluid dynamics data).

This integration aspect is involved in all DT stages, from creating DTs to the use phase (simulations and what-if scenarios), and finally to maintaining the twin.

Current UDTs are mainly based on semantic city models. These models are considered the foundation for data integration. However, if we change this model and extend it, we are performing integration at level 1. On the other hand, when the default model supports the data to be integrated, we are at level 2. However, if we add architecture components to embed the data, we are at level 3. The action of integrating enables model enrichment to create comprehensive DTs of the real world. The creation of DTs needs to be discussed upstream to ensure that models can be maintained. In fact, current implementations use existing models and adapt them in certain ways to fulfil DT requirements. In the DT creation phase, we refer to the creation of city models while emphasizing that these models are developed to form the basis of the UDT. Indeed, we are lacking a generic DT that can be relevant when creating regional or national DTs. At the city level, however, we tend to focus on implementations based on use cases, which leads to the generation of multiple DTs and, by default, affects the integration levels. We have noticed that in the creation phase, we usually use semantic models of current cities, such as Helsinki, Vienna, and Japan, to name a few. These models generally integrate data based on levels 1 or 2 in the creation phase. Even in the use phase, ADEs, or generics (objects or attributes), are deployed. Other implementations make extensive use of level 3 to integrate data, but more precisely during the use phase.

In the present work, we wanted to highlight the data integration levels and offer some conceptual guidelines according to the DT's lifecycle to enhance data integration and interoperability. Data integration enables DT developers to shift from a semantic city model to a UDT. Integration takes place at different levels, and various configurations of levels can be envisaged, notably in UDT, where several types of data are required. For each phase of the DT lifecycle, we integrate data, and for each phase, one or more levels of integration can be applied.

Moving from a city model to a UDT, DT designers need to plan the integration of different types of data. The best level of integration does not exist. In fact, each UDT project has its own appropriate integration level, and this is also guided by the DT lifecycle phases. For instance, in the creation phase, it is important to consider that these models will be used in simulations and urban applications, which in turn generate new data that needs to be reincorporated. In this regard and following the analysis provided in Section Section 5.1, level 3 in the creation phase is still limited to few implementations given that UDT developers tend to use level 1 of integration based on the conceptual data model. This is a valid assumption since the core of Urban Digital Twins are 3D city model that represents the foundation layer of integration from geospatial perspective. However, in the use phase, level 3 is commonly used. To support integration at this level, database-independent architectures are developed.

For the update phase, we are unable to draw any conclusions on the appropriate level of integration, given that in the current implementations, we were unable to identify any proposed mechanisms for updating UDTs. This is also true since almost all current implementations are still in the creation and use phases. However, to overcome this, we would like to set up experiments for future work to understand how the DT parties conceive this UDT update.

While this study helps to provide fundamental insights on the UDT challenges related to data integration, however, it has some limitations that will be tackled in future work. One of the limitations is the selection criteria we have used to filter projects. To illustrate our data integration levels, we examined DT projects in detail, which was not a straightforward process since DT developers do not always describe the way in which they integrate their data and whether they are really considering a generic and standardized integration approach to homogenize their data. Another aspect is that the current initiatives are mainly focused on the creation and use stages of digital twins, yet there is a minimal distinction between the creation and use phases. Furthermore, our framework is conceptual and uses real-world implementation to prove its findings, which raises the issue related to the feasibility of implementing all the integration scenarios previously defined in future work.

In this study, we have focused on the most common datasets, such as geodata and simulation outputs. However, taking into consideration the advancements in Generative AI, such models will in their turn bring new integration challenges. At the current maturity stage of UDT, these models help to address the issues related to the data quality, availability, and completeness by generating comprehensive and realistic data through data enhancement and enrichment. They are generally used in urban studies to produce synthetic data, automate the processing in terms of semantic segmentation, and generate 3D city models.

7. Conclusions

Cities worldwide start investing in developing UDTs as a new digital approach to collecting, integrating, managing, visualizing, and sharing data and models. UDTs bring both potentialities and issues. Data integration is identified as a significant challenge associated with the full implementation of the UDT concept. While data integration has been considered a fundamental issue in 3D GIS for a decade, it is still involved in the development of new concepts such as UDTs, which integrate multiple and heterogeneous data. Indeed, addressing the data integration issue will enrich the existing 3D city models and deliver a mature version of the DTs. In this review paper, we provided the first conceptual data integration insights to create, use, and update an UDT. Three data integration levels (conceptual data model, database, and front-end levels) are defined and illustrated according to the DT lifecycle. For each phase of the lifecycle, the data might be integrated using one or several levels on the same UDT platform. The conceptualization of the data integration levels is a domain- and application-independent classification. That means that the classification proposed in this work might be applied to any UDT implementation. Furthermore, we believe that new data integration levels will appear in the future to handle UDT data and models. Regardless of the advantages and limitations reported for each data integration level, data quality, completeness, and interoperability strongly guide integration effectiveness. Our defined data integration levels will help researchers and practitioners (namely cities and municipalities) that have already an ongoing UDT implementation or those who are willing to embark on this journey towards the creation of urban digital twins to understand and transfer the common practices related to data integration and to discuss upstream the integration frameworks, the level(s) required according to the available data and the user needs. The users can be government, researchers, private sector, and finally citizens.

In future research, we will focus on technically testing the different scenarios, investigating them based on the DT lifecycle, and formulating guidelines to recommend effective data integration levels based on different standards and frameworks. Future steps aim to investigate in more detail each scenario defined in this work and test it in a specific case study using various datasets. We believe that implementing technically the different data integration configurations will help us to determine their relevance, evaluate the robustness and applicability of the generic data integration levels in different use cases, and assess the most effective level of integration. To reach our intended goal, we should establish a technological framework and provide solutions in terms of data models, databases, and visualization tools, which will accordingly assist the development of the UDT. An exploratory survey will be designed in the future to get insights from UDT practitioners about their thoughts regarding these data integration levels and if their initiatives follow any of these integration levels. The survey focus will also be on evaluating the differences of data integration in the maintenance phase. This will bridge the current gap in the literature in terms of how integration is carried out in the maintenance stage.

Notes

1. https://github.com/tum-gis/SDDI-CKAN-Docker?tab=readme-ov-file.
2. https://www.ogc.org/standard/CityGML/.
3. https://cesium.com/platform/cesiumjs/.
4. https://developers.arcgis.com/javascript/latest/.
5. https://www.esri.com/en-us/arcgis/products/arcgis-cityengine/overview.
6. https://www.esri.com/en-us/arcgis/products/arcgis-geobim/overview.
7. https://vc.systems/en/.
8. https://docs.ogc.org/per/22-041.html#_0998bc4e-baa2-49b0-901c-ae2572a41b8f.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Belgian National Funds for Scientific Research FNRS. This research is part of the project GIS 3.0 that demonstrates the convergence of Geographic Information Systems and Web 3.0: Semantic Web techniques, object-oriented prototype languages (JavaScript, JSON,) and document-oriented NoSQL databases. The research project (PDR) is funded by the Belgian National Funds for Scientific Research FNRS_2019_SIG3.0_PDR/OL T.0024.20.

ORCID

Imane Jeddoub http://orcid.org/0000-0003-0632-1406
Gilles-Antoine Nys http://orcid.org/0000-0001-5580-5680
Rafika Hajji http://orcid.org/0000-0002-1291-397X
Roland Billen http://orcid.org/0000-0002-3101-8057

Author's contribution

Imane Jeddoub: Conceptualization, Writing – original draft, Writing – review & editing. **Gilles-Antoine Nys**: Conceptualization, Review and editing. **Rafika Hajji**: Conceptualization, Supervision, Review & editing. **Roland Billen**: Conceptualization, Supervision, Review & editing. All authors have read and agreed to the published version of the manuscript.

References

Abdalla, R. 2016. "Geospatial Data Integration." In *Introduction to Geospatial Information and Communication Technology (GeoICT)*, edited by R. Abdalla, 105–124. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-33603-9_6.

Agugiaro, G., J. Benner, P. Cipriano, and R. Nouvel. 2018. "The Energy Application Domain Extension for CityGML: Enhancing Interoperability for Urban Energy Simulations." *Open Geospatial Data, Software and Standards* 3 (1): 2. https://doi.org/10.1186/s40965-018-0042-y.

Alva, P., F. Biljecki, and R. Stouffs. 2022. "Use Cases for District-Scale Urban Digital Twins." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-4/W4-2022:5–12. https://doi.org/10.5194/isprs-archives-XLVIII-4-W4-2022-5-2022.

Billen, R., A. Cutting-Decelle, O. Marina, J.-P. Almeida, M. Caglioni, G. Falquet, T. Leduc, et al. 2014. "3D City Models and Urban Information: Current Issues and Perspectives." https://doi.org/10.1051/TU0801/201400001.

Boumhidi, K., G.-A. Nys, and R. Hajji. 2024. *Integrating Dynamic Data with 3D City Models via CityJSON Extension*, 745–758. https://doi.org/10.1007/978-3-031-43699-4_45.

Chaturvedi, K. 2021. *Integration and Management of Time-dependentProperties with Semantic 3D City Models*. Phd thesis., Chair of Geoin-formatics, Technical University of Munich. https://www.researchgate.net/publication/364581384.

Chaturvedi, K., B. Willenborg, M. Sindram, and T. H. Kolbe. 2017. "Solar Potential Analysis And Integration Of The Time-dependent simulation Results For Semantic 3d City Models Using Dynamizers." *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences* IV-4/W5:25–32. https://doi.org/10.5194/isprs-annals-IV-4-W5-25-2017.

Chaturvedi, K., Z. Yao, and T. H. Kolbe. 2019. "Integrated Management And Visualization Of Static And Dynamic Properties Of Semantic 3d City Models." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4/W17:7–14. https://doi.org/10.5194/isprs-archives-XLII-4-W17-7-2019.

Dembski, F., U. Wössner, M. Letzgus, M. Ruddat, and C. Yamu. 2020. "Urban Digital Twins for Smart Cities and Citizens: The Case Study of Herrenberg, Germany." *Sustainability* 12 (6): 2307. https://doi.org/10.3390/su12062307.

Diakite, A. A., L. Ng, J. Barton, M. Rigby, K. Williams, S. Barr, and S. Zlatanova. 2022. "LIveable City Digital Twin: A Pilot Project For The City Of Liverpool (NSW, Australia)." *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences* X-4/W2-2022:45–52. https://doi.org/10.5194/isprs-annals-X-4-W2-2022-45-2022.

DImitrov, H., and D. Petrova-Antonova. 2021. "3D City Model as a First Step Towards Digital Twin of Sofia City." *Presented at the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* XLIII-B4-2021:23–30. https://doi.org/10.5194/isprs-archives-XLIII-B4-2021-23-2021.

Döllner, J., T. H. Kolbe, F. Liecke, T. Sgouros, and K. Teichmann. n.d. "The Virtual 3d City Model of Berlin-Managing, Integrating, and Communicating Complex Urban Information." Proceedings of the 25th International Symposium on Urban Data ManagementUDMS 2006, 1–12. Aalborg, Denmark.

Fricke, A., J. Döllner, H. Asche, and F. Isaac. 2023. "Optimising a Virtual 3D City Model for Citizen-Based Urban Development in East Jerusalem." In *Computational Science and Its Applications – ICCSA 2023 Workshops, Lecture Notes in*

Computer Science, edited by O. Gervasi, B. Murgante, A. M. A. C. Rocha, C. Garau, F. Scorza, Y. Karaca, and C. M. Torre, 521–535. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37114-1_36.

Haraguchi, M., T. Funahashi, and F. Biljecki. 2024. "Assessing Governance Implications of City Digital Twin Technology: A Maturity Model Approach." *Technological Forecasting & Social Change* 204:123409. https://doi.org/10.1016/j.techfore.2024.123409.

Hu, Y. 2017. *Spatial Data Infrastructures. Geographic Information Science & Technology Body of Knowledge 2017*. https://doi.org/10.22224/gistbok/2017.2.1.

IEC. 2021. "City Information Modelling and Urban Digital Twins | IEC [WWW Document]." https://www.iec.ch/basecamp/city-information-modelling-and-urban-digital-twins .

Janowicz, K., S. Schade, A. Bröring, C. Keßler, P. Maué, and C. Stasch. 2010. "Semantic Enablement for Spatial Data Infrastructures." *Transactions in GIS* 14 (2): 111–129. https://doi.org/10.1111/j.1467-9671.2010.01186.x.

Jeddoub, I., G.-A. Nys, R. Hajji, and R. Billen. 2023. "Digital Twins for Cities: Analyzing the Gap Between Concepts and Current Implementations with a Specific Focus on Data Integration." *International Journal of Applied Earth Observation and Geoinformation* 122:103440. https://doi.org/10.1016/j.jag.2023.103440.

Ketzler, B., V. Naserentin, F. Latino, C. Zangelidis, L. Thuvander, and A. Logg. 2020. "Digital Twins for Cities: A State of the Art Review." *Built Environment* 46 (4): 547–573. https://doi.org/10.2148/benv.46.4.547.

Khawte, S. S., M. N. Koeva, C. M. Gevaert, S. Oude Elberink, and A. A. Pedro. 2022. "Digital Twin Creation for Slums in BBrazil Based on UAV data." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-4/W4-2022:75–81. https://doi.org/10.5194/isprs-archives-XLVIII-4-W4-2022-75-2022.

Kiourtis, A., A. Mavrogiorgou, and D. Kyriazis. 2024. "A Cross-Sector Data Space for Correlating Environmental Risks with Human Health." In *Information Systems*, edited by M. Papadaki, M. Themistocleous, K. Al Marri, and M. Al Zarouni, 234–247. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56478-9_17.

Kleftakis, S., A. Mavrogiorgou, K. Mavrogiorgos, A. Kiourtis, and D. Kyriazis. 2022. "Digital Twin in Healthcare Through the Eyes of the Vitruvian Man." In *Innovation in Medicine and Healthcare*, edited by Y.-W. Chen, S. Tanaka, R. J. Howlett, and L. C. Jain, 75–85. Singapore: Springer Nature. https://doi.org/10.1007/978-981-19-3440-7_7.

Knezevic, M., A. Donaubauer, M. Moshrefzadeh, and T. H. Kolbe. 2022. "Managing urban digital twins with an extended catalog Service." *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences* X-4/W3-2022:119–126. https://doi.org/10.5194/isprs-annals-X-4-W3-2022-119-2022.

Kolbe, D. T. H., M. Moshrefzadeh, K. Chaturvedi, and D. A. Donaubauer. n.d. August 2020. *Updated Version*.

Kolbe, T. H., and A. Donaubauer. 2021. "Semantic 3D City Modeling and BIM." In *Urban Informatics, the Urban Book Series*, edited by W. Shi, M. F. Goodchild, M. Batty, M.-P. Kwan, and A. Zhang, 609–636. Singapore: Springer. https://doi.org/10.1007/978-981-15-8983-6_34.

Kotsev, A., M. Minghini, R. Tomas, V. Cetl, and M. Lutz. 2020. "From Spatial Data Infrastructures to Data Spaces—A Technological Perspective on the Evolution of European SDIs." *ISPRS International Journal of Geo-Information* 9 (3): 176. https://doi.org/10.3390/ijgi9030176.

Kutzner, T., K. Chaturvedi, and T. H. Kolbe. 2020. "CityGML 3.0: New Functions Open Up New Applications." *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science* 88 (1): 43–61. https://doi.org/10.1007/s41064-020-00095-z.

La Guardia, M., and M. Koeva. 2023. "Towards Digital Twinning on the Web: Heterogeneous 3D Data Fusion Based on Open-Source Structure." *Remote Sensing* 15 (3): 721. https://doi.org/10.3390/rs15030721.

La Guardia, M., M. Koeva, F. D'Ippolito, and S. Karam. 2022. "3D DAta Integration For Web Based Open Source WebGL Interactive Visualisation." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-4-W4-2022:89–94. https://doi.org/10.5194/isprs-archives-XLVIII-4-W4-2022-89-2022.

Lehner, H., S. L. Kordasch, C. Glatz, and G. Agugiaro. 2024. "Digital geoTwin: A CityGML-Based Data Model for the Virtual Replica of the City of Vienna." In *Recent Advances in 3D Geoinformation Science, Lecture Notes in Geoinformation and Cartography*, edited by T. H. Kolbe, A. Donaubauer, and C. Beil, 517–541. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43699-4_32.

Lei, B., P. Janssen, J. Stoter, and F. Biljecki. 2023. "Challenges of Urban Digital Twins: A Systematic Review and a Delphi Expert Survey." *Automation in Construction* 147:104716. https://doi.org/10.1016/j.autcon.2022.104716.

Lei, B., Y. Su, and F. Biljecki. 2024. "Humans As Sensors in Urban Digital Twins." In *Recent Advances in 3D Geoinformation Science. 3DGeoInfo 2023. Lecture Notes in Geoinformation and Cartography*, edited by T. H. Kolbe, A. Donaubauer, and C. Beil Cham: Springer. https://doi.org/10.1007/978-3-031-43699-4_42.

Li, W., S. Zlatanova, A. A. Diakite, M. Aleksandrov, and J. Yan. 2020. "Towards Integrating Heterogeneous Data: A Spatial DBMS Solution from a CRC-LCL Project in Australia." *ISPRS International Journal of Geo-Information* 9 (2): 63. https://doi.org/10.3390/ijgi9020063.

Li, W., S. Zlatanova, J. J. Yan, A. Diakite, and M. Aleksandrov. 2019. "A Geo-database Solution For The Management And Analysis Of Building Model With Multi-source DAta Fusion." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-4-W20:55–63. https://doi.org/10.5194/isprs-archives-XLII-4-W20-55-2019.

Noardo, F. 2022. "Multisource Spatial Data Integration for Use Cases Applications." *Transactions in GIS 26* 26 (7): 2874–2913. https://doi.org/10.1111/tgis.12987.

Noardo, F., K. Arroyo Ohori, F. Biljecki, C. Ellul, L. Harrie, T. Krijnen, M. Kokla, and J. Stoter. 2020. "The ISPRS-EUROSDR Geobim Benchmark 2019." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B5-2020:227–233. https://doi.org/10.5194/isprs-archives-XLIII-B5-2020-227-2020.

Padsala, R., S. Banerjee, T. Belayneh, T. Santhanavanich, and V. Coors. 2023. "Conceptualizing I3S Encoding for OGC CityGML 3.0 Building Model." https://doi.org/10.1145/3611314.3615920.

Padsala, R., E. Gebetsroither-Geringer, J. Peters-Anders, and V. Coors. 2021. "Inception of Harmonising Data Silos And Urban Simulation Tools Using 3d City Models For

Sustainable Management Of The Urban Food Water And Energy Resources." *Presented at the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* VIII-4/W1-2021:81–88. https://doi.org/10.5194/isprs-annals-VIII-4-W1-2021-81-2021.

Rossknecht, M., and E. Airaksinen. 2020. "Concept and Evaluation of Heating Demand Prediction Based on 3D City Models and the CityGML Energy ADE—Case Study Helsinki." *ISPRS International Journal of Geo-Information* 9 (10): 602. https://doi.org/10.3390/ijgi9100602.

Santhanavanich, T., and V. Coors. 2021. "CityThings: An Integration of the Dynamic Sensor Data to the 3D City Model." *Environment & Planning B: Urban Analytics & City Science* 48 (3): 417–432. https://doi.org/10.1177/2399808320983000.

Santhanavanich, T., R. Padsala, M. Rossknecht, S. Dabirian, M. Saad, U. Eicker, and V. Coors. 2023. "ENabling Interoperability Of Urban Building Energy Data Based On Ogc Api Standards And Citygml 3d City Models." *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences* X-1/W1-2023:97–105. https://doi.org/10.5194/isprs-annals-X-1-W1-2023-97-2023.

Santhanavanich, T., R. Padsala, P. Würstle, and V. Coors. 2022. "THe Spatial Data Infrastructure Of An Urban Digital Twin In The Building Energy Domain Using Ogc Standards." *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences* X-4/W2-2022:249–256. https://doi.org/10.5194/isprs-annals-X-4-W2-2022-249-2022.

Sreemathy, J., R. Brindha, M. Selva Nagalakshmi, N. Suvekha, N. Karthick Ragul, and M. Praveennandha. 2021. "Overview of ETL Tools and Talend-Data Integration." 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, 1650–1654. India. https://doi.org/10.1109/ICACCS51430.2021.9441984.

Stoter, J. E., G. A. K. Arroyo Ohori, and F. Noardo. 2021. "Digital Twins: A Comprehensive Solution or Hopeful Vision? GIM International: The Worldwide Magazine for Geomatics 2021."

Stoter, J. E., G. A. Ohori, B. Dukai, A. Labetski, K. Kavisha, S. Vitalis, and H. Ledoux. 2020. "State of the Art in 3D City Modelling: Six Challenges Facing 3D Data as a Platform. GIM International: The Worldwide Magazine for Geomatics 34."

Weil, C., S. E. Bibri, R. Longchamp, F. Golay, and A. Alahi. 2023. "A Systemic Review of Urban Digital Twin Challenges, and Perspectives for Sustainable Smart Cities." *Sustainable Cities and Society* 99:104862. https://doi.org/10.1016/j.scs.2023.104862.

Würstle, P., R. Padsala, T. Santhanavanich, and V. Coors. 2022. "VIability Testing Of Game Engine Usage For Visualization Of 3d Geospatial Data With Ogc Standards." *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences* X-4/W2-2022:281–288. https://doi.org/10.5194/isprs-annals-X-4-W2-2022-281-2022.

Würstle, P., T. Santhanavanich, R. Padsala, and V. Coors. 2020. *The Conception of an Urban Energy Dashboard Using 3D City Models*, 523–527. https://doi.org/10.1145/3396851.3402650.

Xu, H., F. Omitaomu, S. Sabri, S. Zlatanova, X. Li, and Y. Song. 2024. "Leveraging Generative AI for Urban Digital Twins: A Scoping Review on the Autonomous Generation of Urban Data, Scenarios, Designs, and 3D City Models for Smart City Advancement."

Zheng, X., J. Lu, and D. Kiritsis. 2022. "The Emergence of Cognitive Digital Twin: Vision, Challenges and Opportunities." *International Journal of Production Research* 60 (24): 7610–7632. https://doi.org/10.1080/00207543.2021.2014591.