Multisource spatial data integration for use cases applications

Running title: Multisource spatial data integration

Francesca Noardo[1]

3D geoinformation group, Urbanism Department, Delft University of Technology, Delft, The Netherlands

[1] Corresponding author - f.noardo@tudelft.nl, francesca@noardo.eu ORCID:0000-0003-2269-5336

Acknowledgements

This study has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Leading Fellow Postdoc programme, grant agreement no. 707404, project "Multisource spatial data integration for smart city applications". I'd like to thank the 3D geoinformation group at TU Delft, who gave great support and inspiration to my research.

Multisource spatial data integration for use cases applications

The reuse and integration of data give big opportunities, supported by the F.A.I.R. data principles. Seamless data integration from heterogenous sources has been interest of the geospatial community for long time. However, 3D city models, BIM and information supporting smart cities present higher semantic and geometrical complexity, which pose new challenges, never tackled in a comprehensive methodology. Building on previous theories and studies, this paper proposes an overarching workflow and framework for multisource (geo)spatial data integration. It starts from the definition of use case-based requirements for the integrated data, guides the analysis of integrability of the involved datasets, suggesting actions to harmonise them, until data merging and validation. It is finally tested and exemplified on a case study. This approach allows the development of consistent, well-documented and inclusive data integration workflows, for the sake of use cases automation in various geospatial domains and the production of Interoperable and Reusable data.

Keywords: data integration; interoperability; harmonization; GeoBIM; metadata.

1. Introduction

The integration of spatial information from heterogeneous sources has been of interest in the geomatics community for a long time. For the professional use of spatial data, the effective interaction of multi-source data is extremely useful. Recently, the new opportunities given by the technological developments in acquisition and surveying techniques, which allow the generation of high amounts of data, as well as the technologies to share the data, mainly through the web (e.g., open data, linked data), opened new perspectives towards the reuse of existing data for further use cases than the ones for which they were originally collected. Similarly, the Findable, Accessible, Interoperable and Reusable (F.A.I.R.)[1] data principles support a new sharing economy for data.

However, although many initiatives have been developed towards interoperability, such as standardization actions, ontology-related research (Kavouras, Kokla, 2007), GeoBIM research (for the integration of geoinformation and Building Information Models), a comprehensive methodology for multi-source data integration has not been proposed yet.

Mohammadi et al. (2006) identify the aspects to be considered for the data integration as: institutional, policy, legal and social, besides technical. This paper is focused on the technical side of the integration.

According to Kavouras and Kokla (2007), to build an integrated view of heterogeneous systems requires: identifying the heterogeneities; analysing importance and priorities; solving them through a systematic strategy. Wiemann and Bernard (2016) define the steps for data integration, called ‘data fusion’ as: data search and retrieval; data enhancement; harmonization; relation measurement; feature mapping; resolving; data provision. They propose an interesting approach based on linked data. Mohammadi et al. (2010) proposed a methodology and a tool to facilitate spatial data integration within spatial data infrastructures, considering the technical and non-technical issues. However, the increased complexity of data available nowadays (3D, deep and complex data structures) presents new challenges from the technical point of view that need to be tackled to achieve an effective integration.

Existing efforts, often consider mainly the semantic and structural aspects for data integration (e.g. Lenzerini, 2002). For example, Kavouras and Kokla (2007) outline

[1] https://www.howtofair.dk Accessed 06/06/2022

relevant methodologies for ontologies integration, which are partially re-used, adapted and eventually referred within this paper. However, their focus is on semantics and structure, while it is important for this study to consider the complexity of geometry as a separate issue. Furthermore, data from practice hardly reach the complexity considered by Kavouras and Kokla (2007).

Other cases report methods to merge the geometric information from multiple sources or sensors (‘data fusion’) (e.g., Dong et al., 2018; Ramos, Remondino, 2015; Zhu, Donia, 2013; Ahn et al., 2020). Data fusion techniques are developed to integrate (big) data by means of different criteria and methods intended to automate the integration, i.e.: data-level, feature-level, decision-level fusion (Zang et al., 2022; Yin et al., 2021). More complete database integration was also studied, but still for 2D GIS (Devogele et al., 1998, Uitermark et al., 2005).

Standardization efforts have been intended to solve the interoperability, and consequently integration, issues. However, their development is still little aligned with their implementation in software and adoption in practice (Noardo et al., 2020, 2021a,b,c), which makes the original aim of standardization still an open problem (Section 1.1).

In this paper, the features of spatial data are analysed in detail, including the characteristics defining the complex 3D information systems, such as Building Information Models (BIM) and 3D city models. The needs of the processing towards harmonization of the data is defined accordingly, and some relevant available methods to perform such processing are mentioned as initial guidance. Those phases are inserted in an overall workflow guiding from the choice of the input datasets until the final merging and validation of the harmonised data. A critical starting step to initiate a successful integration process is the definition of the requirements for the finally obtained integrated data, to support the intended use cases applications. A concrete and well-defined scope and use of the data (including software and procedural details) is the preferable way to success and allows validation and testing.

While most of existing methodologies focus on few aspects of data integration, sometimes neglecting the features of data as provided in practice, a pragmatic workflow is proposed in this work to support overall data integration. The adopted approach strongly relates on the definition of requirements for the integrated data, and decomposes the integration into sub-issues, to be analysed and tackled in detail.

The methodology combines methods common in ontology engineering and data fusion, with experiences in 3D information systems integration and GeoBIM, as described in Section 2. The results are explained starting from the proposed integration workflow (Section 3) and the parameters and specific features to be considered in the integration (Section 3.1). After some suggestions for data requirements definition (Section 3.2) and data retrieval (Section 3.3), a rubric to analyse each feature or parameter of the input data, assessing their potential for integration, is proposed in Sections 3.4 and 3.5. In Section 3.6, a range of methods are mentioned to tackle the processing pointed out by such analysis and assessment. Finally, Section 3.7 presents methods for fusing the harmonised data, and validation. The proposed methodology is exemplified in a case study (Section 4). The discussion (Section 5) follows, including the potential automation of the proposed workflow (Section 5.1) and an analysis of metadata standards (Section 5.2) possibly facilitating the preliminary integrability assessment.

1.1 Open standards and related issues

To support interoperability and integration, Open standards are published for different domains. For the representation of city and wider portions of land, CityGML[2] was published by the Open Geospatial Consortium (OGC). OGC also recognised CityJSON[3] (Ledoux et al., 2019) as community standard, intended to improve usability of CityGML, mainly by means of a different implementation. INSPIRE[4] and its proposed data model is a further reference for the representation of city and landscape, prescribed by the European Directive 2007/2/EC aimed at providing a common data infrastructure to support environmental policies across Europe. To represent infrastructure knowledge from a geospatial point of view, the OGC LandInfra[5] was published. In the field of Architecture Engineering Construction and Operations, mainly for buildings, other standards have been developed, such as the Industry Foundation Classes[6], by buildingSMART, or gbXML[7] with the specific scope of representing energy-related features of buildings and constructions, to support analysis.

However, these standards also present disadvantages, like the big efforts for producing compliant data or the reduced flexibility in some cases (Doerr, 2004). At the same time, they often propose very comprehensive schemas, aiming at covering the entire domain, and leave the possibility open to use the model in very different ways, to adapt to different use cases' needs. Although it makes such standards suitable for a large variety of representations, interoperability and integration processes suffer from this, because data become quite unpredictable, even if compliant to the same standard. It is

[2] https://www.ogc.org/standards/citygml [Accessed 1st December 2021]

[3] https://www.cityjson.org [Accessed 1st December 2021]

[4] https://inspire.ec.europa.eu [Accessed 1st December 2021]

[5] https://www.ogc.org/standards/landinfra [Accessed 1st December 2021]

[6] https://www.buildingsmart.org/standards/bsi-standards/industry-foundation-classes/ [Accessed 1st December 2021]

[7] https://www.gbxml.org [Accessed 1st December 2021]

essential to know how the (often ambiguous) models were interpreted, with respect to structure, semantics and geometry, and how the structure was used, e.g., which of the allowed options is used to store some specific information, such as georeferencing (Clemen, Görne, 2019). In addition, extensions and generic entities can be used to extend the prescribed model further, resulting in an even wider possibility to produce standard-complaint conflicting data. Clear definitions and examples are still seldom available to allow a consistent use of the standard data models (including interpretation of classes meaning, attributes and relationships), although the general tendency towards the improvement of such aspects in the current standardisation efforts.

Moreover, when considering the use and implementation into concrete tools and data from practice, they often present issues (Noardo et al., 2020, 2021a,b,c) and are not used consistently enough to provide fully automatically interoperable and integrable data.

Figure 1 shows an example of three 3D city models of Rotterdam (NL). One is the 3D city model developed by the City of Rotterdam[8], the second was generated by the software 3dfier (Ledoux et al., 2021) and the third one is the Basisregistratie Grootschalige Topografie (BGT), the Dutch national topographic map, and is structured according to the IMGeo[9] data model, which is modelled as an Application Domain Extension (ADE) of CityGML v.2 (Van den Brink et al., 2013a). All of them are CityGML-compliant. However, they result in quite different models (Colucci et al., 2020).

[8] https://www.3drotterdam.nl/#/ [Accessed 24th November 2021]

[9] https://www.geonovum.nl/geo-standaarden/bgt-imgeo [Accessed 24th November 2021]

3D Rotterdam	3Dfier	BGT
LoD2	LoD1	LoD0
core:xalAddress		
bldg:Building	bldg:Building	bldg:Building
bldg:yearOfConstruction		
core:creationDate		
bldg:measuredHeight	bldg:measuredHeight	bldg:measuredHeight
building number		DGDT_ID
lowest_building layer		
highest_building layer		
number of building layers		
deviation		
statusOmschr		
typeOmschr		
avineonStatus		
		CLASS
		CLASS_IMG
bldg:BuildingInstallation		
bldg:BuildingPart		
bldg:ClosureSurface		
bldg:GroundSurface		
bldg:OuterCeilingSurface		
bldg:RoofSurface		
bldg:WallSurface		
furnit:CityFurniture		
	genobj:GenericCityObject	genobj:GenericCityObject
		CLASS PHYSICS CF CLASS_IMG
veget:SolitaryVegetationObject		
	trpt:Road	trpt:Road
		CLASS FUNCTION PHYSICS CF CLASS_IMG
	landuse:LandUse	landuse:LandUse
		CLASS PHYSICS CF CLASS_IMG
	water:WaterBody	water:WaterBody
		CLASS
	brid:Bridge	

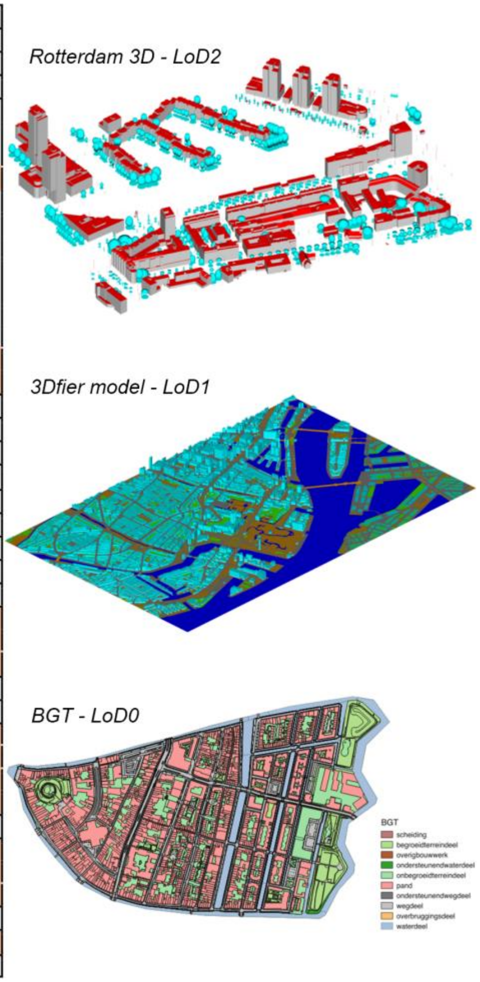


Figure 1. A comparison of entities, attributes (in pink the attributes used as generics) and visualization of three different CityGML-compliant models.

As a consequence, an in-depth analysis is necessary to consider the relevant parameters and characteristics involved in interoperability, explained in Section 3.1, even if the data are declared standard-compliant.

1.2 Interoperability vs data integration

The two concepts of interoperability and integration are often confused. However, even if being strictly related, they have a different meaning.

Kavouras and Kokla (2007) state that ‘interoperability is the ability of systems or products to operate effectively and efficiently in conjunction, on the exchange and reuse of available resources, services, procedures, and information, in order to fulfil the requirements of a specific task'. They add that 'it is not exhausted with integration, but also involves means of intelligent communication such as querying, extraction, transformation etc.'. Moreover, interoperability in a broader governance-related domain is defined as 'the ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems' (EU, 2017). In such a context, four interoperability layers are identified: technical, semantic, organisational and legal interoperability. These are interconnected and include the usual aspects considered for interoperability: technology, regarding information and communication technology systems and software; data; humans, i.e. needed skills, know-how and related general knowledge and practice; institutional practices, i.e. the processes and best practices implemented in everyday life within institutions and practice. The scope of this paper covers the aspect of data, in particular in relation to the so-called 'semantic' interoperability, but has strict relationship and influence on the technical interoperability of data and on the human side of interoperability, concerning data interpretation and description for re-use (grey dotted rectangle in Figure 2).

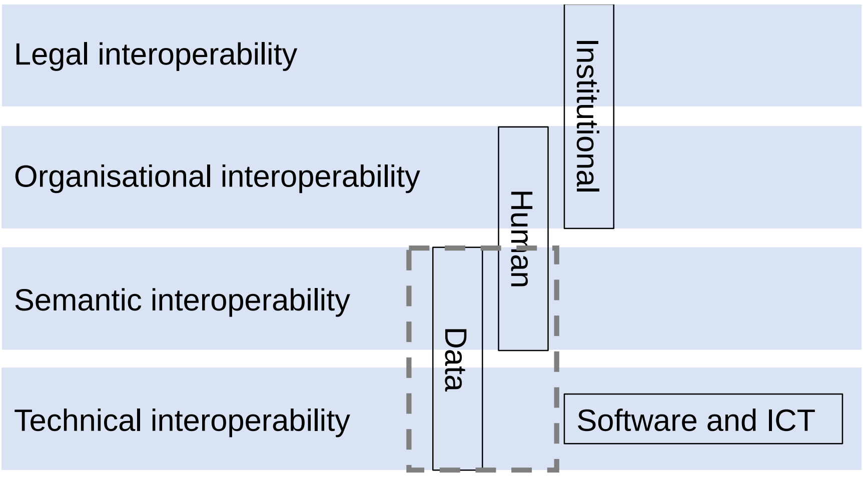

Legal interoperability
Organisational interoperability
Semantic interoperability
Technical interoperability
Institutional
Human
Data
Software and ICT

Figure 2. Interoperability layers from EU (2017) with the aspects of interoperability mapped (data, humans, institution and technology). The grey dotted rectangle identifies the scope of this paper with respect to interoperability.

Interoperability can be considered as a characteristic of single data, allowing their re-usability across systems (e.g. their potential for being consistently imported-exported by software) (Noardo et al. 2021a,b).

Integration is instead the combination or conflation of information from different data sets (Worboys, Duckham, 2004).

Figure 3 depicts what the two concepts entail and how are they related to each other.

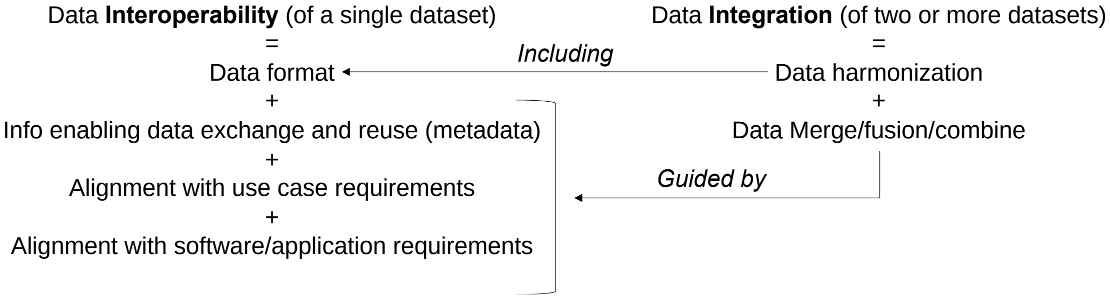

Data **Interoperability** (of a single dataset)
=
Data format
+
Info enabling data exchange and reuse (metadata)
+
Alignment with use case requirements
+
Alignment with software/application requirements
Including
Data **Integration** (of two or more datasets)
=
Data harmonization
+
Data Merge/fusion/combine
Guided by

Figure 3. Data interoperability vs data integration.

This study is particularly focused on data integration.

2. Materials and methods

Reaping the results from recent studies on the integration of geoinformation and BIM (GeoBIM) and past data integration theories and experiences (Noardo, 2020a,b; Ellul, 2020; Biljecki, Tauscher, 2019; Arroyo Ohori et al., 2018; Noardo et al., 2016; Kumar et al., 2019; Ulubay, Altan, 2002; Laurini, 1998; Laurini, Thompson, 1992; Worboys Duckham, 2004; Kavouras, Kokla, 2007), the complex issue of spatial data integration is analysed in its components and a reference workflow is proposed.

Moreover, the relevant set of data parameters and aspects to be considered for the integration is explicated and organised into a framework. It is intended to provide a reference for assessing the integration potential of datasets with respect to destination data requirements, as well as to guide in the harmonization.

Processing methods are then suggested per each case, to transform and convert the input datasets, as necessary to harmonize them with the data requirements. Potential usable methods are reviewed.

The methods to solve each of the steps in the integration workflow are many and need to be chosen according to the kind of processing needed and the kind of data involved. Therefore, in this paper it is not possible to give one recipe to fit all cases, but several methods from literature are proposed in order to overcome the most usual issues. The proposed framework is intended to point out the needs and guide in the process.

The proposed framework is finally validated and tested in a case study, regarding the update of a 3D city model by means of the integration of a BIM model of a newly designed building. The focus of the experiment is not on the processing itself, therefore, many steps were performed manually or by means of existing tools. Other similar cases were proposed, e.g., by Noardo et al. (2016) and van Heerden (2021).

3. The proposed workflow and framework for multisource data integration

A workflow for an effective methodology for data integration is depicted in Figure 4. It is comprehensive of the starting phase, i.e. definition of requirements, until the final phase regarding the update of metadata after data merging. Moreover, the several issues for data integration are considered, including the legal constraints, harmonisation, data merging and validation. Available examples in literature usually focus on a part of it, without considering the other aspects explicitly. Since this paper is intended as a general framework to guide the integration concretely, we need to consider all the steps consistently.

The integration effort for a specific use case is considered. Therefore, the essential starting point, critical in this methodology, is the definition of the requirements for the data to be obtained after the integration (Section 3.2). The parameters to be defined in the data requirements definition, as well as in the following phases, are listed in Section 3.1. In addition, the non-technical features, as defined in Mohammadi et al. (2006), must be planned. Based on such defined requirements, the input datasets must be retrieved in order to cover the defined need for information (Section 3.3). It is necessary to double check that the legal properties of input data are not conflicting with the integrated ones (e.g., copyright, privacy, licences and so on.). In case they conflict, it should be assessed whether input data can be transformed to meet the requirements (possibly through data generalisation, anonymisation, attribute removals etc.). Otherwise, either an adjustment should be done in integrated data requirements (i.e., future data specs), whether possible, or other data must be retrieved. Then, the assessment of data integrability must be performed, according to the framework proposed in Sections 3.4 and 3.5. From that analysis it can be assessed whether it is possible to use the selected input data in the integration, and if the required effort to harmonise them is worth doing it. Otherwise, different data should be selected. If the assessment is positive, the harmonisation actions (enrichment, generalization, conversion) must be chosen and applied for each considered aspect (Section 3.6). Data fusion will then allow obtaining the integrated data set (Section 3.7). Final steps are the validation of such data set against the defined data requirements and the update of metadata to keep track of the applied processing.

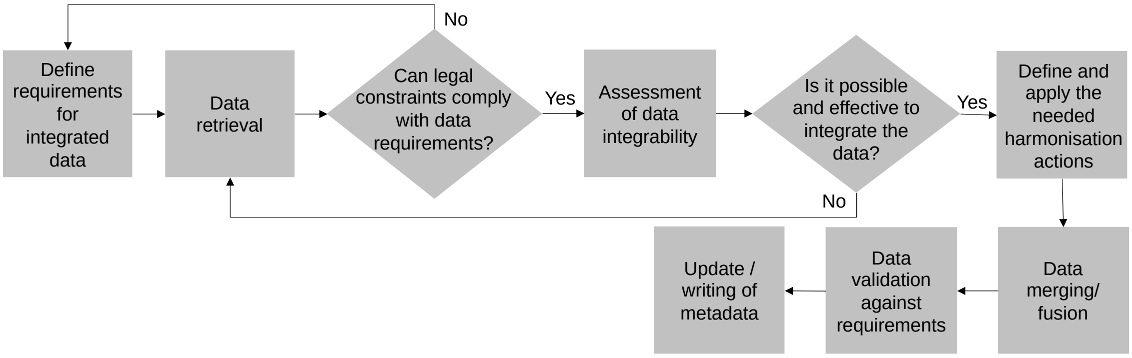

Define requirements for integrated data
Data retrieval
Can legal constraints comply with data requirements?
No
Yes
Assessment of data integrability
Is it possible and effective to integrate the data?
Yes
No
Define and apply the needed harmonisation actions
Data merging/ fusion
Data validation against requirements
Update / writing of metadata

Figure 4. A workflow for suitable data integration

3.1 Relevant data features and parameters.

In this Section, the data characteristics are explained that should be always explicitly prescribed by data requirements in case of data acquiring and modelling, or harvesting, described in metadata and considered in data integration efforts (Figure 5).

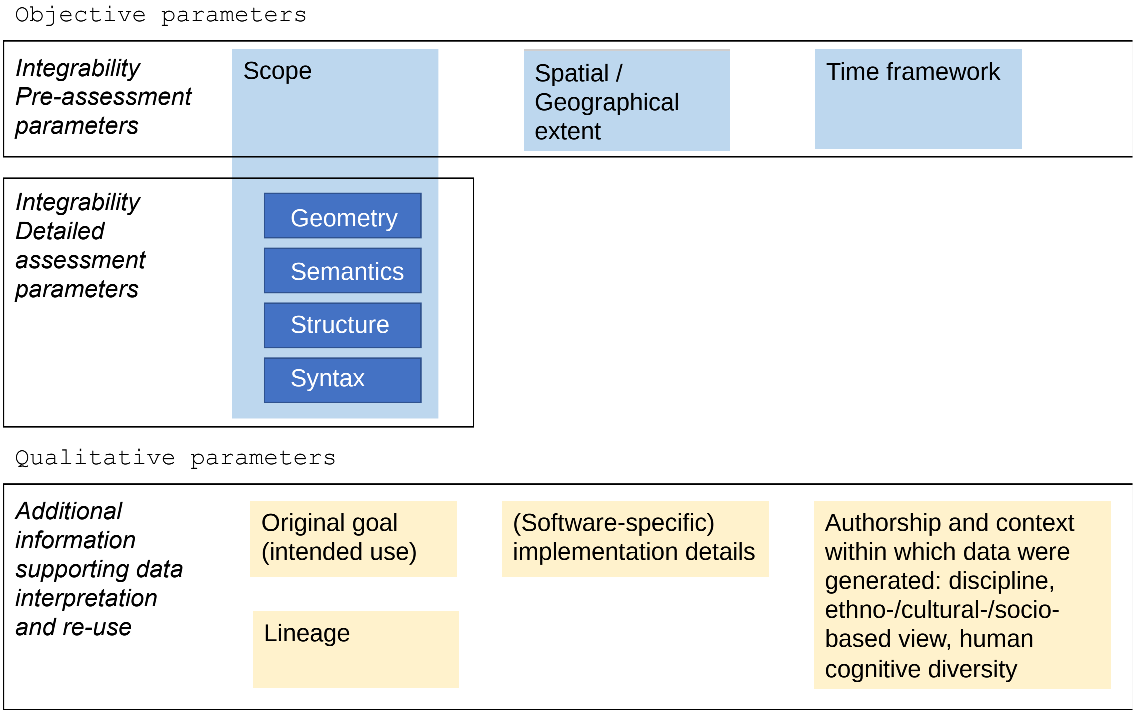

Objective parameters
Integrability Pre-assessment parameters
Scope
Spatial / Geographical extent
Time framework
Integrability Detailed assessment parameters
Geometry
Semantics
Structure
Syntax
Qualitative parameters
Additional information supporting data interpretation and re-use
Original goal (intended use)
(Software-specific) implementation details
Authorship and context within which data were generated: discipline, ethno-/cultural-/socio-based view, human cognitive diversity
Lineage

Figure 5. Synthesis schema of the data parameters relevant for data integration, as considered in this paper.

Resources on spatial data management and integration distinguish between ‘*semantic’* level (i.e. difference in conceptualization and definition - including terms used, specific meaning and classifications); *‘structural’* or *‘schematic’* level (i.e. the conceptual model or schema structuring the data – relations between entities and attributes, relationships, hierarchies); and *‘syntactical’* level (i.e. the format of the data). Semantics and structure are very much interrelated, and one cannot be considered without the other. In fact, they are often considered together simply as ‘semantics’. However, they are treated separately here, according to other authoritative sources in literature (Doerr, 2004; Kavouras, Kokla 2007; Worboys, Duckham, 2004).

What is generally missing in the literature dealing with data integration theory is the geometry level. Often, this is treated as part of the semantic content of the data, such as in the ontologies-related field (Kavouras, Kokla, 2007). In other studies, geometry is the main focus of integration, like in data fusion and 2D GIS integration literature, but the others are neglected. In this paper, the *‘geometric’* level is considered as a separated aspect of the data, since it presents specificities which need to be tackled by means of specific methods. Overlooking the geometry could cause serious issues to data integration, reducing the potential for such data to be used within automatic tools. Obtaining a consistent integration including of geometry, instead, supports a powerful reuse of data for various use cases, such as exploiting the BIM or GIS software for (geo)spatial data analysis and smart cities support.

In addition, more general properties of the data must be considered, related to their contents and intended use, in order to make a preliminary assessment about their usability for the designed goal. These are the geographical extent, time frame and scope. The *geographical extent* is the real location of the objects represented by the data. It is defined by a specific spatial extent located with respect to the Earth surface, which could be 2D (planar location) or 3D (considering heights and z values). The *temporal frame* indicates the time period in which data were acquired or updated. The *scope* of the data is the definition of the part of reality to be represented and the intended use for which the conceptualization was designed. It determines differences with respect to: coverage and detail, granularity (e.g. building and building elements vs all city objects); classification perspectives and consequent relations (e.g. building door as part of the internal distribution system of the construction vs building door as address); semantics, due to the kind of partition of reality, i.e. specific intended meaning of a term or concept, with respect to synonymy, homonymy and different meanings for the same term (e.g. slab intended as structural element or slab intended as all the package dividing the storeys from each other) (Kavouras, Kokla, 2007); geometry. Therefore, considering the general scope of data in the initial assessment could be meaningful even before analysing the previously mentioned parameters, that specify the scope in detail (geometry, semantics, structure and syntax).

Geographical extent, temporal framework and scope can be documented as objective information. Further, it is relevant to take into account some more qualitative background of the data, which has an impact on modelling and implementation choices, to properly (re)use the data and avoid mistakes in their interpretation.

The *original goal* of the data (intended use) and the specific use case requirements, for which they are produced, influence the data themselves (‘perspective’ in Kavouras, Kokla, 2007), affecting the modelling (geometry, semantics, structure) and storage (syntax, data format).

The *lineage* of the data determines their final characteristics that must be known by the data users (accuracy, objects represented etc.) (e.g., Thapa, Bossler, 1992; Biljecki et al., 2015; Lunetta et al., 1991). It implies modelling method, including the original sources of information possibly processed for the modelling phase (e.g. previous maps, original survey, acquisition and measurement methods, measure processing, storage methods, point clouds, photogrammetric plotting, etc.), the criteria used and the choices made for modelling the final data (represented objects, used LoD, generalization methods and any other pre- or post-processing). Many times, it is sufficient to know the resulting characteristics, but in the most complex cases, documenting the process of production of a dataset can help understand it and use it as properly as possible, avoiding the propagation or generation of errors or misuse and misinterpretation of the data.

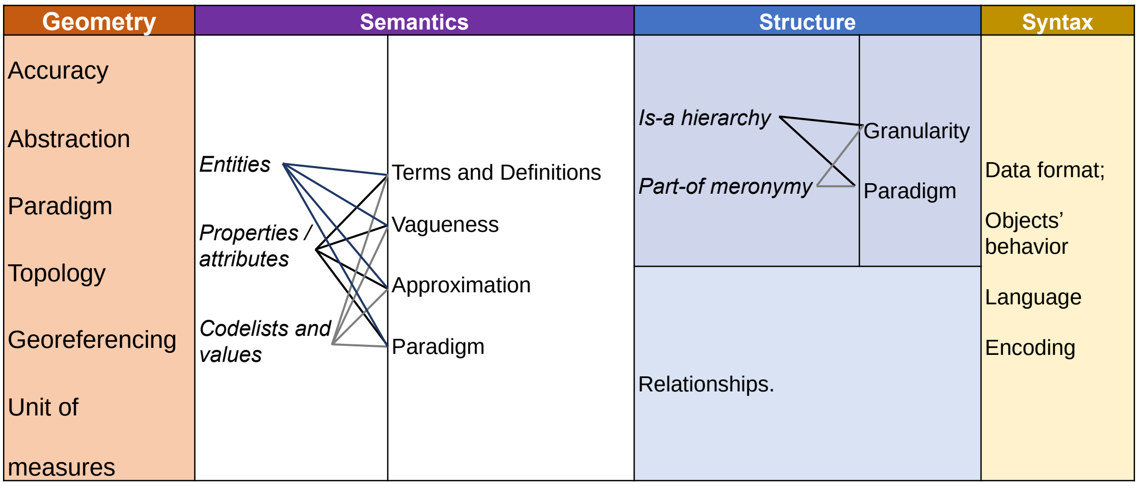
Finally, *(software-specific) implementation details* – the software that will use the data (or for which/with which the data were modelled) – must be known, since they also determine choices in the use, selection and storage of information within the datasets.

Besides these, Kavouras and Kokla (2007) add, among the causes of taxonomic diversity: discipline (field into which data are designed and generated); ethno-/cultural-/socio-based view (nuances in the concept meaning and interpretation of a domain by different cultures or societies, as well as the local geographical terms used); human cognitive diversity (different individuals perceive and conceptualise a domain differently). Therefore, also knowing the *authorship* of data and the *context* within which they were generated can help manipulating them correctly. Such diversities are not only relevant for taxonomy, but for the whole bunch of choices made when producing the data, therefore affecting also the geometric modelling, the format chosen and so on. Integration must be aware of all the diversities involved.

Metadata should describe as many as possible of the listed parameters, which are critical to assess the data correctly and speed up the integration, guaranteeing a higher quality of the data resulting from the integration.

Table 1 summarizes the geometric, semantic, structural and syntactical parameters involved in data definition relevant for integration, which are explained in detail in annex 1.

Table 1. Synthesis of parameters for data integration potential assessment



Geometry	Semantics		Structure		Syntax
Accuracy Abstraction Paradigm Topology Georeferencing Unit of measures	*Entities* *Properties / attributes* *Codelists and values*	Terms and Definitions Vagueness Approximation Paradigm	*Is-a hierarchy* *Part-of meronymy*	Granularity Paradigm	Data format; Objects' behavior Language Encoding
			Relationships.		

3.1.1 A reflection on (Open) standard-compliant data.

Structure, as well as semantics and geometry, can be declared compliant to (Open) standard data models. However, this does not per se solve the harmonization-related issues, since many data from practice can implement or interpret the standard data models differently (Section 1.1). In addition, a few variables must be taken into account.

First, the *version* of the standard must be checked. Although, in theory, backwards compatibility should be guaranteed, new versions may dismiss part of the schema or introduce new classes, attributes and codelists. New specifications may constrain the use of the standard. Consequently, data generated according to new specifications may result different from data compliant to previous versions. For example, in the CityGML version n.3[10], entirely new modules are introduced with respect to version n.2 (OGC, 2012)[11], such as ‘Construction’, to describe construction details, as well as other modules related to the performance of the model, like ‘Versioning’ and ‘Dynamizer’. Other changes are made in the core module, introducing the concept of ‘FeatureType’ (i.e., abstractions of real-world phenomena that have an identity) disjoint from ‘ObjectType’ (i.e. objects that have an identity but are not features). It makes the general conceptualisation quite different from the previous version, and compliant models can therefore differ over different versions of the same standard accordingly. The concept of space was mainly modelled by ‘Room’ class in CityGML 2, while there is a different and more detailed conceptualisation foreseen in version 3 (AbstractLogicalSpace, AbstractPhysicalSpace, occupied or unoccupied space and so on).

Another example, among many others, is the storage of georeferencing information in IFC v.2x3 with respect to v.4, where the storage of more complete information is enabled (Clemen, Görne, 2019).

Besides that, often not the entire standard data model is used in datasets, but only a *profile*, that is a part of the entire model, according to the needs of applications. It represents the actual data model used by the data and it is therefore relevant to outline it

[10] https://docs.ogc.org/is/20-010/20-010.html [Accessed 24th November 2021]

[11] https://www.ogc.org/standards/citygml [Accessed 24th November 2021]

explicitly. The description of the specific interpretation and use of the model is very relevant to enable consistent use and integration.

Similarly, *extensions* of the standard data models are used to enhance their representation scope for specific applications, by means of foreseen mechanisms, such as the Application Domain Extension (ADE) in CityGML (Van den Brink et al., 2013b). In case official extensions are used, they can be considered similar to a reference data model themselves, therefore, the version and used profile must be verified and compared as well.

Moreover, *generics* (classes, attributes and relationships) are foreseen in the standard data models (‘generics’ module in CityGML, *ifcProxies* in IFC). They provide a structure for objects not covered by any other class, attribute or relationship in the standard conceptual model. Although the recommendation is to use them only if an appropriate structure is not provided by the rest of the schemas, in data from practice they are very often used in place of existing entities. Therefore, even if standard-compliant, many data follow a customized data model (Noardo et al., 2021b; Colucci et al., 2020).

Such variations of standard data models (profile, extension, use of generic classes) should be documented in proper metadata and documentation associated to the dataset, including, preferably, the formal encoding and the parameters described in this Section (3.1). It would enable the automation of the mapping and conversion of compliant datasets. However, most of datasets coming from practice do not provide a proper explicit documentation and it makes it necessary to analyse the data manually.

Additional variability could come from different interpretations of the same data model. A translation or conversion or even enrichment / new acquisition can be necessary whether the interpretation of the data model is too far from data requirements.

Adopting Open standards correctly, even if using different profiles and extensions, would give anyway the advantage of speeding-up the mapping to support the following integrability assessments.

3.2 Define requirements for the integrated data

As when modelling new data sets, to obtain proper data for the desired application and use case, data requirements for the data to be integrated must be defined. Some standards propose guidelines to define data requirements properly, for example in the building and civil engineering works domain, the concept of Level Of Information Need is established by the ISO 19650-1:2018[12] for information stored in BIM. Meanwhile, buildingSMART defines the Information Delivery Specification (IDS) standard to define the exchange requirements in a computer interpretable format, to define the Level of Information Need and allowing data validation[13]. In the geospatial domain, data requirements must also be defined, according to the use case for which they are intended (Malinowski, Zimányi, 2006).

It is essential to define such data requirements as well for the information resulting from integration, considering the mentioned standards and covering all the parameters listed in Section 3.1. They will later guide all the following steps of the integration (data retrieval, information selection, harmonization and processing and data merging/fusion/combining).

In data requirements, tolerance thresholds can also be set to establish the admitted discrepancy with respect to the defined parameters.

[12] https://www.iso.org/obp/ui/#iso:std:iso:19650:-1:ed-1:v1:en [Accessed 24th November 2021]

[13] https://technical.buildingsmart.org/projects/information-delivery-specification-ids/ [Accessed 24th November 2021]

3.3 Data retrieval

In this paper we will not analyse the issues related to data retrieval (findability, usability, licenses, costs and so on). However, this does represent a further aspect to be considered in the integration. The legal and policy constraints must be checked against the possibility to use the data for integration and the data requirements (including foreseen use and publication of the data and so on).

If any technical (e.g., the necessary information is not present or not suitable) or non-technical (e.g., costs, licence, etc.) issue is found, a new data retrieval phase is necessary.

Data retrieval include both existing datasets and possible new acquisitions, whether it is not possible to find any suitable resource.

3.4 Pre-assessment of data integrability

A preliminary assessment about the effectiveness of data sets integration can be performed by comparing initial information about the geographical extent, time frame and general scope (Figure 6). Scope will be later specified as the geometry, semantics and structural characteristics described in Section 3.5. The preliminary information could be found in good metadata, otherwise it will be necessary to retrieve it, whether possible, by inspecting the data or annexed documentation.

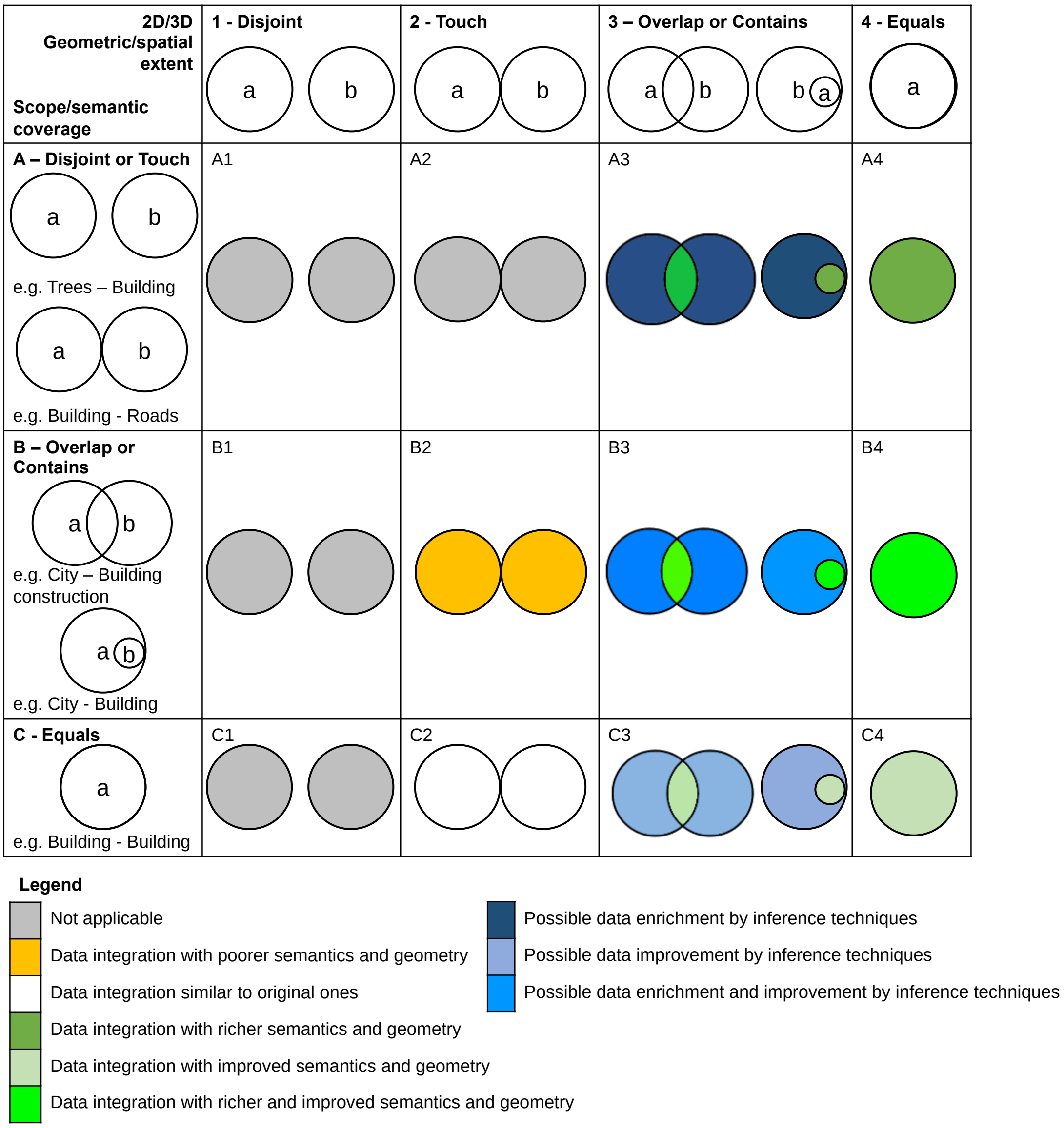

2D/3D Geometric/spatial extent
Scope/semantic coverage
1 - Disjoint
2 - Touch
3 – Overlap or Contains
4 - Equals
a
b
A – Disjoint or Touch
e.g. Trees – Building
e.g. Building - Roads
A1
A2
A3
A4
B – Overlap or Contains
e.g. City – Building construction
e.g. City - Building
B1
B2
B3
B4
C - Equals
e.g. Building - Building
C1
C2
C3
C4
Legend
Not applicable
Data integration with poorer semantics and geometry
Data integration similar to original ones
Data integration with richer semantics and geometry
Data integration with improved semantics and geometry
Data integration with richer and improved semantics and geometry
Possible data enrichment by inference techniques
Possible data improvement by inference techniques
Possible data enrichment and improvement by inference techniques

Figure 6. Overview of the possible data integration based on the spatial and logical relationships of geographical extent and scope/semantic coverage. In the legend, ‘richer’ information indicates a higher quantity of information (e.g. both buildings and trees, while I had only buildings before); ‘improved’ information indicates a higher quality of the information (e.g. I select the most accurate information about the same object between the two datasets).

Examples of datasets covering different scopes and semantic coverage are: A – any 3D city model (a) and weather data related to sensors represented in a map (b); B – a 3D city model or map (a) and a BIM (b) or a 2D digital map (a) and a CityGML model containing only buildings, as there are many examples (b); C – two CityGML models (e.g. Figure 1).

In cases A1, A2, B1, C1, in Figure 4, it does not make sense to integrate the data, unless by trying with complex inferences and machine learning or data mining processing (e.g. Perumal et al., 2015; Wang, 2017; Sheeren, 2004). In A3 and A4 (for example, one dataset about buildings and one about roads covering different geographical extensions which, at least partially, overlap) the overlapping area can be integrated to obtain a richer dataset (i.e. a dataset including both roads and buildings).

In the B2 case (e.g. a GIS and a BIM located in areas bordering each other), data can be integrated to obtain a more extended dataset with the overlapping semantic coverage of the two datasets. In the example of GIS and BIM data, I could obtain a new, more extended dataset probably including less information than the original sources, since I should, for example, select only part of the information in the GIS, to keep only buildings, and generalize the information in the BIM to obtain the same representation present in the GIS. In B3 and B4 (e.g. a GIS and a BIM whose geographical extensions overlap), the overlapping area can be integrated to obtain a richer dataset (i.e. with more information) by integrating the information related to different semantic coverage. For example, I can have the information about the building in BIM, such as materials and windows, plus the information about the building in GIS, such as function and address or owner, plus the context elements present in GIS, such as trees and roads. On the other hand, the overlapping semantic coverage of the two datasets (e.g. building) can be used to improve the quality of the data through the integration, by mapping and comparing the data, and selecting the best or most suitable version of the information. A further example could be a map produced for topographic representation of the land and a map produced for running a risk analysis on a specific area, or maps produced by different institutions (e.g. national or regional authority and a municipality).

In C2 (e.g. the topographical maps of two bordering municipalities), data can be integrated to obtain a similar dataset on a wider extent. Mapping and information quality comparison can help checking consistency on the final data (slightly improved or decreased depending on details of the harmonisation processing).

In C3 and C4 (e.g. the topographical maps provided by the national or regional authority and by a municipality about overlapping areas), in the overlapping areas, the two datasets should be quite similar. Still, they could differ for the geometric representation or differences in data quality, which would make the integration useful to improve the data through mapping and information quality comparison (I can choose to keep the most accurate information, for example). Moreover, the data could be enriched in case they use different profiles of the same schemas or extensions and generics which complement each other.

In A3, B3 and C3 cases, inferences techniques can be assessed for enriching the data in the non-overlapping area based on the integration processing on overlapping part. I could analyze the relationships of data in the overlapping area and try to reproduce them in the remaining part starting from the present information. For example, I can detect a pattern between the year of construction of buildings (possibly present in one dataset) and the materials used (possibly present in the other dataset) and infer the missing information for the part of the datasets which are not overlapping.

B1 and C1 could be similar to the B2 and C2 cases, respectively. However, the two datasets could be successfully harmonised and converted to a common format, but the final step of data merging/fusion or combining would not make sense. It must be decided based on use cases whether the effort is useful. For example, it could make sense to harmonise the data whether this allows running the same analysis, using the same tools on the two separate areas (e.g. a flood analysis tool), and possibly compare the results.

Additional discussion would be necessary about the use of data referring to different time frameworks. In this case, whether the spatial extents overlap at least in part (cases 3 and 4 in Figure 4), and the scope and kind of representation as well (cases B and C in Figure 4), the data could be integrated for the overlapping part, at least, if setting the priorities and criteria for merging based on the data requirements. Optimal case would be using data referred to close time frameworks.

3.5 Detailed assessment of data integrability

It is necessary to run a preliminary data quality assessment with respect to the defined data requirements, to answer the question: ‘Is it appropriate and effective to integrate the data?’. For some of the parameters considered, there are quality thresholds deciding the suitability of the dataset (e.g. the minimum accuracy), while in other (most of) cases, the data can be improved through pre-processing in order to reach the needed quality.

In this case as well, metadata and annexed specifications should be analysed first, to speed up the process whether they report suitable and reliable information.

Tables 2,4,5,6 define a rubric to assess the level of integrability of datasets, i.e., their compatibility with respect to the data requirements prescribed and the needed pre-processing. Based on the data requirements, availability of alternative data, effort required and processing options, the user will assess the suitability of the data to be involved in the integration or whether a different choice of data (including new acquisition and processing) should be preferred. In the tables, scores are given based on the scale: 0 – the data cannot be used for integration; 1 – the data must be pre-processed through complex data mining/machine learning/ data enrichment processing; 2 – the data must be pre-processed to be generalised properly; 3 – a conversion must be applied; 4 – the data can be used as they are.

3.5.1 Geometry integrability assessment

Table 2 explains the criteria to assess geometry-related parameters. Regarding *'accuracy'*, on the condition that both datasets respect the minimum accuracy required, there are no studies demonstrating specific issues when integrating datasets having different accuracies. Although it is preferable the two datasets having similar accuracy, it is possible to consider valid a dataset coming from two datasets having different accuracies as well. However, future studies will investigate the issue in more detail, to identify possible challenges and thresholds in the difference of accuracy between the datasets involved in the integration.

A minimum condition for georeferencing also applies. Georeferencing information, with the minimum accuracy required by data requirements, can be later converted or inferred, but an indication of georeferencing parameters, reference points or at least a qualitative description about the data location is necessary.

Table 2. Level of integrability based on the geometry-related parameters.

	0 – not usable	1 - enrichment	2 - generalization	3 - conversion	4 – as is
Accuracy	< data requirements' accuracy	-	-	-	≥ data requirements' accuracy

Abstraction	The needed objects are not represented	Represented objects need more details for a higher LoD or higher-resolution	Geometries must be generalised to a lower LoD or lower resolution	-	Same than prescribed in data requirements
Paradigm	-	-	-	Geometries must be converted to a different representation	Same than prescribed in data requirements
Topology	Topological relationships stored as prescribed by requirements cannot be seen/inferred	Topological relationships can be inferred	Topological relationships can be generalised	Topological relationship must be expressed in a different way.	topological relationships stored as prescribed by the data requirements
Georeferencing	No information about location	Little, vague, inaccurate information (e.g. address or adescription when specific coordinates are required)	-	Information is in a different CRS or stored differently than prescribed in requirements	CRS, accuracy and storage as prescribed in data requirements

Units o.m.	-	-	-	The model must be scaled and/or converted to a different unit	Same than prescribed in data requirements

3.5.2 Semantics integrability assessment

Semantics consist in the concepts expressed by objects names, as well as by the terms defining attributes and composing code lists. Those should be described within proper definitions, removing possible ambiguity from interpretations.

Considering both terms and definitions while mapping the concepts and objects represented in two different data sets allows a higher reliability in the similarity assessment (Table 3), as well as considering their features and semantic neighbourhood and context (Rodríguez, Egenhofer, 2003; Kavouras, Kokla, 2007). The reliability of the mapping still increases when considering attributes and instances as well. In case of geographical datasets, the spatial dimension of instances can also be used to establish correspondence between features (Rodríguez, Egenhofer, 2003).

Table 3. Different combination of Term (T) and Definition (D) cases (Kavouras, Kokla, 2007)

	1) T1=T2	2) T1≠T2
a) D1=D2	a1) Equivalence	a2) Synonymy
b) D1>D2	b1) Further investigation required	b2) IS-A
c) D1∩D2	c1) Overlap	c2) Overlap
d) D1≠D2	d1) Homonymy	d2) Disjointness

doi:10.20944/preprints202112.0286.v2

Terminological-conceptual conflicts can be: ‘confounding’ – information items having deceptively the same meaning but are actually differing (e.g. a ‘tie-beam’ was interpreted as ‘IfcBeam’ in an example reported by Noardo et al., 2022); ‘scaling’ – from the use of different reference systems and scale; ‘naming’ – from using different terms (homonyms and synonyms) (Kavouras, Kokla, 2007; Wahe et al., 2001).

The possibility to convert between different semantics paradigms, including between different calculation methods and filling criteria for attributes, should be assessed for the specific cases: e.g. calculation could be based on further data or values that must be known etc. Condition is that the semantic paradigm, filling criteria and methods used are well documented within metadata.

The semantic paradigm must be compatible with the data requirements. For example, Noardo et al. (2016) reports differences in filling the roads classification in Italian and French digital maps, being ‘paved’/ ‘unpaved’ in the Italian maps and classified according to a hierarchy of functions in the French maps. In this case, it is hard to infer or calculate the values from the available data, and a third source of information is likely necessary.

Table 4. Level of integrability based on the semantics-related parameters. Each parameter must be taken into account for: entities; properties and attributes; codelists values and attribute values.

	0 – not usable	1 - enrichment	2 -generalization	3 - conversion	4 – as is

Terms	Cases d) (Table 3)	Cases b), definition of data is larger than definition in requirement; Cases c) partial information is present in data	Cases b), definition of data is narrower than definition in data requirement	Case a2) – synonymy	Case a1) – equivalence
Vagueness	> data requirements' vagueness	Generic terms (e.g. low/ medium/ high) can be mapped to more specific data (e.g. value intervals)	Data can be generalized (e.g. specific intervals can be mapped to low/medium/ high terms).	-	≤ data requirements' vagueness
Approximation	The needed objects are not represented	Lower approximation needed by data requirements	Higher approximation needed by data requirements	-	Same than prescribed in data requirements
Parad	Incompatible	-	-	Compatible	Same

3.5.3 Structure integrability assessment

Table 5 defines the criteria to assess the integrability of data based on the structural parameters.

Table 5. Level of integrability based on the structure-related parameters.

		0 – not usable	1 -enrichment	2-generalization	3 - conversion	4 – as is
For Is-a hierarchies and Part-of	granularity	The needed objects are not. represented	Higher specification and deeper hierarchy needed by requirements	Lower specification. shallower hierarchy needed by requirements	-	Same than prescribed by data requirements
	Paradigm	Incompatible	-	-	Compatible	Same
Relationships		Absent and not inferable	Relationships can be inferred	Relationships can be generalized	Relationships (names, encoding, criteria) can be converted	Same than prescribed by. data requirements

3.5.4 Syntax integrability assessment

Table 6 defines the criteria to assess the integrability of data based on the syntactical parameters.

Table 6. Level of integrability based on the syntactic-related parameters.

	0 – not usable	1 - enrichment	2- generalization	3 - conversion	4 – as is
Data format	-	-	-	Data formats are different	Same data format

Objects' behaviour	Needed objects' behaviour are not present	Objects' behaviours must be added	Objects' behaviours must be removed	Objects' behaviours must be converted	Same objects' behaviours
Language	-	-	-	Different language	Same language
Encoding	-	-	-	Different encoding	Same encoding

3.5.4 The final overall assessment

The integrability potential of the dataset can be roughly measured by summing up the scores related to all the parameters considered and reported by data requirements. If any of the parameters has scored 0, the integration cannot be performed and the process should be blocked. The maximum integrability rate is the total number of parameters, calculated as in equation: *number of involved parameters*4* (i.e., all the parameters scored 4 and the data set can be merged as is), while the minimum should be the total number of parameters (all the parameters scored at least 1, i.e., the dataset can be used after an enrichment that is possible). However, this is only a rough assessment, and the processing to harmonise the dataset with respect to each parameter needs to be considered singularly.

Moreover, it should be noticed that the assessment can regard only the part of the data set which is intended to be used in the integration. On the other hand, some of the parameters can be irrelevant for the data requirements (e.g., there are no attributes or code lists). Checking that all the information prescribed in data requirements is covered by the datasets involved in the integration must be done separately, in the initial phase of data retrieval and/or later, during the validation step.

3.6 Define the needed harmonisation actions.

Once the data sets involved are assessed as suitable for the integration (integrability scores 1 to 3 as defined in Section 3.5), a pre-processing must harmonise their characteristics with the ones indicated by data requirements (to reach integrability score 4). In this section, the needed actions are listed, together with possible methods, referring to each parameter and integrability score case. The Subsection 3.6.1 introduces the semantic and structure mapping, as defined within the ontology engineering field. It is preliminary to understand the suggested approach in processing and harmonizing the data in this paper.

3.6.1 Ontologies and data models mapping and integration

The ‘schema mapping’, or ‘schema matching’ (implying also semantics) is the definition of an automated transformation of each instance of a data structure A into an instance of a data structure B that preserves the intended meaning of the original information (Doerr, 2004). The preservation of the intended meaning must be ultimately judged by the application domain expert.

Doerr (2001) defines the principles for mapping thesaurus terms by means of concept-based mapping. Mapping and ontology integration techniques are the tools necessary to solve most of the inhomogeneities in semantics, in ‘Terms’ choice, and in the structure. For this reason, although each parameter is considered as a separated issue in this paper, for not neglecting any of them, a mapping between the classification followed by the input data and the one defined by the data requirements is the preliminary step to guide the following processing to harmonise the semantics and structure features.

The techniques and methodologies for mapping different data models or ontologies consider schematic and semantic differences, including syntactic and semiotic/pragmatic heterogeneities in some cases (Kavouras, Kokla, 2007). There are several approaches to integrate ontologies, many of which are based on inter-ontology mapping and alignments between multiple ontologies (Wahe et al., 2001).

According to Kavouras and Kokla (2007), ontology integration approaches can be defined according to three dimensions: D1) the possible change/alteration/distortion caused; D2) the number of ontologies resulting from the integration process; D3) the use of a target ontology in the integration process. For the scope of this paper, an integration involving possible change/alteration/distortion is admitted (D1), one only data model will result from the integration (D2) and the ontology or data model defined in the data requirements will be used as target (D3). The destination data model or ontology could correspond with one of the involved data sets schemas or being a third one defined by data requirements. A hybrid approach (Wahe et al., 2001) is the recommended one (Figure 7).

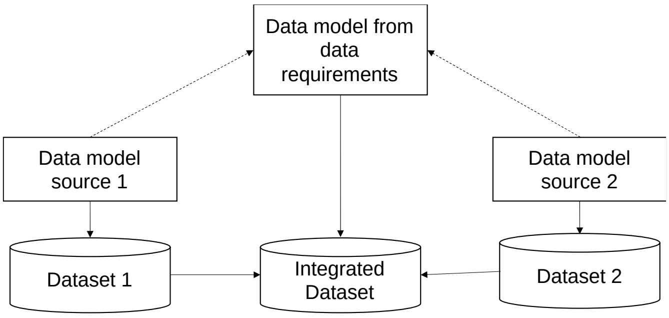


Figure 7. Recommended integration approach in this paper.

Ontology integration consists of (the iteration of) the following steps (McGuinness et

doi:10.20944/preprints202112.0286.v2

al., 2000; Klein, 2001):

(1) Matching, i.e., the procedure that compares concepts and matches those that are more similar in meaning according to a given context (Kavouras, Kokla, 2007; Klein, 2001), or find where the two conceptualizations overlap (McGuinness et al., 2000; Klein, 2001);

(2) Alignment to be generated accordingly (Osman et al., 2021), i.e., mapping of correspondences (equivalence or subsumption relation) between concepts in two or more ontologies into mutual agreement (Kavouras, Kokla, 2007) making them consistent and coherent (Klein, 2001), to overcome syntactic (representation format), terminological (naming differences), conceptual (coverage, granularity and perspective) and semiotic/pragmatic (how the ontology is interpreted or used by communities with respect to a context) heterogeneities (Wahe et al., 2001).

(3) Merging or integrating ontologies is defined as the creation of a new ontology from two or more existing ontologies with overlapping parts, which can be either virtual or physical (Klein, 2001). Osman et al. (2021) define and analyse in more detail the types of processes that can be involved in the above-mentioned cases, as a useful reference for semantic integration processes.

(4) Check the consistency, coherency and non-redundancy of the result.

For the aims of this paper, the generation of a new ontology (step 3) is not of interest, but rather, the mapping will be useful to determine the selection, interpretation and processing to convert the semantics and structure of the input datasets into the ones defined within the data requirements.

3.6.2 Harmonisation of the geometry

With reference to Table 2, the needed harmonisation actions with respect to each geometry parameter are described in this section.

Accuracy. There is no needed pre-processing to adjust the accuracy, since it is a simple threshold value (either sufficient or not). In case the accuracy of the dataset is not homogeneous, it should be verified that the relevant contained information, which is necessary for the integration respect the accuracy threshold established.

Depending on the use cases, it can be assessed case by case whether it is necessary to add acquisition uncertainty to possibly too accurate data (Biljecki et al., 2015) can be used. However, there is no evidence, at the moment, showing that excessively accurate data could bring issues to the integration process.

Similarly, there is no proved research so far showing a reliable method to improve data accuracy without new acquisitions. Some studies propose to improve accuracy by referring to a reliable existing dataset (Noskov, Doytsher, 2017). It could be considered case by case if a similar approach could be applied and what does it entail in 3D.

Abstraction level. Either enrichment or generalization can be necessary.

1 – enrichment) Enrichment is the process of adding missing details to the data in order to increase the amount of information contained. It could use simple processing derived from information contained in the data themselves. For example, the building footprints could be extruded until the height values stored as attribute in the same dataset. Another case foresees the use of external sources to get the necessary information to generate higher levels of detail models. For example, to extrude building

footprints using digital terrain models and heights point clouds as references (Ledoux et al., 2021). More advanced techniques can consider machine learning or other kinds of inferences to complete the unknown information (Biljecki, Dehbi, 2019; Park, Guldmann, 2019)

2 – generalization) Generalization can be applied in case the source data are too detailed. Depending on the kind of data, it means resampling the raster data, reducing detail in the vector data. Studies on generalization are many, some of which are reviewed by Geiger et al. (2015). For 3D city models, for example, it implies extracting lower Levels of Detail from higher LoD representations. Some examples are given by Diakité et al. (2014), Guercke, Brenner (2009), Baig, Abdul-Rahman (2013a,b).

Another case is the generalization of BIMs into 3D city models representations, which imply both a change in the representation paradigm and storage of geometry, and a generalization (Geiger et al., 2015; Donkers et al., 2016; Sun et al., 2019).

In the case of generalization of 3D information systems, including 3D city models, BIM and similar ones, two operations must be considered, i.e., feature extraction, selecting the objects and features to be considered to compute the generalized geometry, and the generalization itself (Guercke, Brenner, 2009). The phase of feature extraction can be based either on geometry (e.g., topological relationships, distance-based or bounding box criteria) or semantics (e.g. including or excluding specific classes of objects, or selecting objects based on attributes – such as the ‘*isExternal’* attribute in IFC).

Geometric paradigm. The harmonization of the geometric paradigm is done via conversion.

3 – conversion) In this case the geometries must be converted into a different representation, as defined by data requirements, for example by means of Extract-Transform-and-Load (ETL) tools (Noardo et al., 2020).

Topology. Data requirements should specify the kind of topological information needed and the kind of storage of such information. Moreover, a validation procedure should be indicated or developed to check that topology is correctly used both in the storage of geometries and in the reciprocal objects' relationships.

1 – enrichment) Data mining, machine learning and inferences techniques (Krijnen et al., 2020) can be used to infer and store a richer topology. Manual techniques could be considered as well in some cases.

2 – generalization) Topological relationships can be generalised to comply with more abstract representations (e.g. Egenhofer et al., 1994).

3 – conversion) The kind of storage of such relationships (being them implicit or explicitly stored in the models) must also be maintained and made compliant to the data requirements (Diakité et al, 2014; Vitalis et al., 2019, Jun, 2019, Zhao, Mbachu, 2019; Salheb et al., 2020).

Georeferencing. It implies the Coordinate Reference System (CRS) for planar coordinates and heights, accuracy of the georeferencing and kind of storage of georeferencing parameters.

1 – enrichment) Techniques to infer the correct georeferencing from too vague data (e.g. the address) can be considered. Options include comparing the model to a different spatial representation (e.g. a BIM model to the parcel where it is supposed to lie, the representation of the model in its context in non-interoperable format, such as PDFs) or a manual positioning based on qualitative information and description. Other

inference techniques can be assessed in order to automate the processing (e.g. Hiebel at al., 2017).

3 – conversion) Re-project to the coordinate and height system prescribed by the data requirements. Techniques can vary based on the kind of data being georeferenced (Noardo et al., 2016; Uggla, Horemuz, 2018; Jaud et al., 2020).

Unit of measure. A simple conversion can harmonize the unit of measure used.

3 – conversion) The geometries should be scaled to the unit of measure needed by the data requirements. ETL tools as well as other modelling software allow this.

3.6.3 Harmonization of the semantics

With reference to Table 4, the needed harmonisation actions with respect to each geometry parameter are described in this section.

Terms and definitions. For the three cases of enrichment (1), generalization (2) and conversion (3), a mapping (Section 3.6.1) must be applied, after having measured and analysed similarity.

Mapping can be done automatically, semiautomatically or manually. Machine learning techniques can also be used (e.g. Doan et al., 2004).

Vagueness. As for the geometric accuracy case, the semantic data should be no vaguer than what admitted by data requirements. It is not possible to enrich the data, because, even if adding more detail starting from the data themselves, the original vagueness would be propagated without reaching the objective.

2-generalization) In case of more accurate data, it is not necessary to consider any processing. In some cases, it is possible that a vaguer value is necessary, for example, a classification ‘high’, ‘medium’, ‘low’, rather than the exact measurement. In such cases, a processing should be applied (and tracked in metadata) to compute and classify the values.

Approximation. As for the case of geometric abstraction level, either enrichment or generalization can be necessary.

1 – enrichment) In case of too vague data, it could be possible to enrich semantics using several techniques, such as ontology-based inferences and machine learning techniques (e.g. Dou et al., 2015; Lüscher et al., 2007; Xue et al., 2021; Bloch, Sacks, 2018; Werbrouck et al., 2020). Attention should be paid for the vagueness value not to be affected by this processing.

2 – generalization) Whether more generalised entities are necessary, superclasses can be used (either from the classification used by the data or mapping the data to other classifications, for example adopting a different perspective). The same techniques for data enrichment could be used, with the different objective of detecting superclasses, in case of differences in the semantic paradigm.

Semantic paradigm. As for geometry, a conversion can harmonize the paradigm.

3 – conversion) The criteria used in defining the conceptualization and filling the attributes have consequences on the resulting meaning of the entities and attribute values and need to be made homogeneous. Whether they are filled with a completely different perspective, objects and attributes must be recalculated or adapted. A transformation, defined through a mapping to a different conceptualization, must be applied, i.e., changing the semantics slightly (possibly also changing the representation) to make it suitable for purposes other than the original one (Klein, 2001). To make a simple example, the mapping of IFC classes to CityGML classes would produce the

doi:10.20944/preprints202112.0286.v2

conversion of IfcRoof to bldg:roofSurface and IfcWall with attribute 'External' to bldg:wallSurface. The concepts are slightly different in the two models, although indicating similar objects, since IFC is intended for construction purposes and CityGML for city analysis goals. Such a distortion needs to be documented and tracked (Kavouras, Kokla, 2007). Moreover, the units of measurements for attributes' values must be converted to comply to the data requirements prescriptions.

Language. A language translation (3 – conversion) must be applied according to the need of application as reported by data requirements. Multilingual thesauri can be used whether available (Doerr, 2001). The buildingSMART Data Dictionary[14] is an example.

Encoding. A conversion (3) is necessary in this case as well.

3.6.4 Harmonization of the structure

With reference to Table 5, the needed harmonisation actions with respect to each geometry parameter are described in this section.

Granularity. Enrichment must be applied in case more detail is needed in the conceptualization, or generalization, in case higher level concepts are necessary.

1 – enrichment) After the mapping, inferences techniques should be applied to specify the objects to a finer granularity (e.g., some 'constructions' will become 'buildings').

2 – generalization) After the mapping, generalization techniques can be applied to generalise the object's conceptualizations, for example, by using superclasses in the

[14] http://bsdd.buildingsmart.org [Accessed 30th November, 2021]

reference classification (e.g. ‘buildings’ and ‘infrastructures’ will become ‘constructions’).

Semantic paradigm. After the mapping, the representation must be expressed according to the semantic paradigm defined in data requirements (3 – conversion).

Relationships. They can be enriched, generalized or simply converted.

1 – enrichment) Relationships between objects can be inferred through the previously mentioned inference techniques, whether the needed information is stored within the dataset. For example, I can detect the apartments represented in an IFC file by starting from the mutual relationships (either semantic or geometric) of the building elements and store the result back in the IFC file according to the proper entity.

2 – generalization) The mapping can support the generalization of relationships, guided by the destination conceptualization. For example, I may need the hierarchical relationship between ‘building part’ and ‘city object’ while the relationship between ‘building part’ and ‘building’ is not of interest for me. I can therefore remove it from the dataset, or better translate them to a relationship with ‘city object’.

3 – conversion) The mapping must regard also the relationships between entities and the input dataset must be converted accordingly. This is the case, for example, in which the relationship is the same but is called differently, such as ‘lives in’ in one dataset and ‘is resident’ in another one.

3.6.5 Harmonization of the syntax

With reference to Table 6, the needed harmonisation actions with respect to each geometry parameter are described in this section.

doi:10.20944/preprints202112.0286.v2

Data format. Encoding language and the representation formalism, including version, must be considered.

3 – conversion) ETL tools, which can also be embedded into GIS tools or other models' exporters, can generally apply the conversion to the desired data formats (Noardo et al., 2020). In ontology engineering field it is called 'translation': changing the representation formalism of an ontology while preserving the semantics (Klein, 2001; KK, 2007).

Objects' behaviour. The mapping will be the reference tool to guide this processing as well in the three cases of enrichment (1), generalization (2) and/or conversion (3).

3.7 Final data fusion and validation

Although in some studies the term 'data fusion' is used to indicate the overall integration, here it is intended to represent specifically the merging of geometrical data overlapping on the same extent or bordering, by resolving the remaining conflicts after the harmonization, which are due to differences in the objects represented by both the sources. For example, discrepancies in DTM heights on certain areas, or in the shape or presence of buildings, and so on.

First, priorities should be decided to choose the most reliable data to be maintained in the integrated data after data fusion. Such priorities should be (1) time (most recent source should be trusted in case of discrepancies); (2) quality (most accurate, less vague source should be preferred, as well as the closest source to data requirements prescriptions); (3) interest (the source representing the objects of interest, whether these are not present in both). Those criteria should be reassessed based on the specific data requirements.

Data fusion is a common process for remote sensing applications, and is usually performed by spatial statistics applications (Nguyen et al., 2012; Zhang, 2010; Ghamisi et al., 2019). Other data fusion examples exist for 2D maps, mainly for data update goals (Duckham, Worboys, 2005; Chen et al., 2013; Devogele, 2002) or for harmonising cross-border maps (Ledoux, Arroyo Ohori, 2017; Noardo et al., 2016). Furthermore, other studies take into account the fusion of complex information and multi-sensor 3D data (Wallgrün, Dylla, 2010; Dong et al., 2018; Ramos, Remondino, 2015; Zhu, Donia, 2013; Ahn et al., 2020).

More sophisticated processing could include the editing of 3D geometries or addition of 3D details to obtain a richer dataset from the merging of two different representations. This should foresee the mapping of features or objects, possible (parametric) modelling phases or other kinds of merging, depending on the manipulated objects and kind of representation. Further studies, not performed so far, will be necessary to investigate the issue in more detail. For example, this could be useful to add specific building details to an already well-formed 3D city model.

The methodology to perform the final merging can be chosen among different options, based on the kind of data considered and the remaining discrepancies. In the final integrated data set the two origin datasets must be no more recognizable, as much as possible, therefore holes and discrepancies in heights must be smoothed. A maximum tolerance in such discrepancy can be considered as the geometric accuracy established by the data requirements.

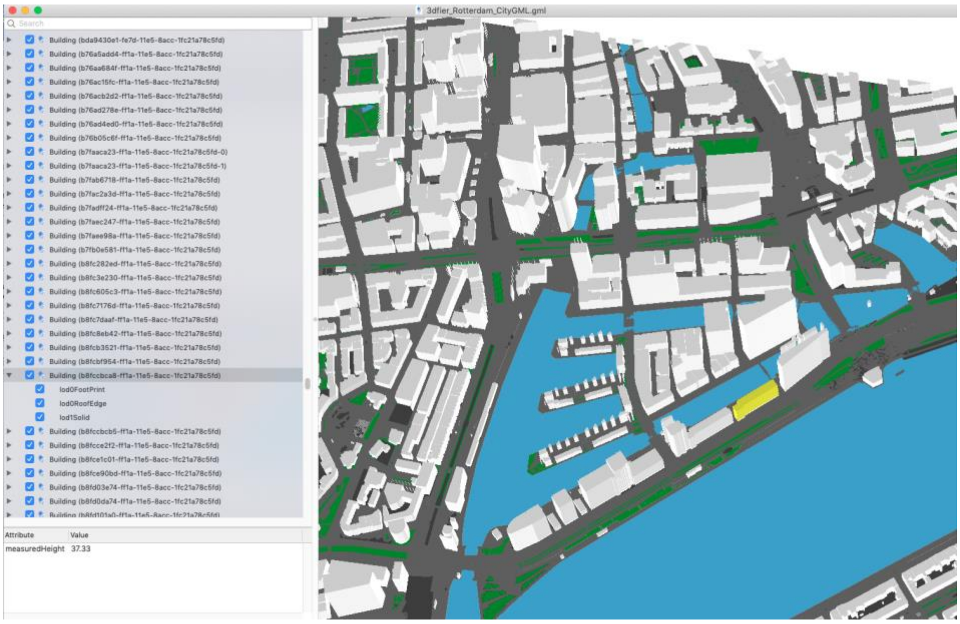
The integrated data obtained should be finally validated against the data requirements. Although it is not covered in detail in this paper, validation is an essential phase of the integration, since it allows the assessment of the integration success and outlines whether the obtained data can be effectively used for the defined use case or not.

In case data requirements were expressed in a formal, machine-readable, language, an automatic validator could be programmed, which is able to read the customised data requirements and check the data against them. However, at present it would be hard to automate the full validation process. Even if, most likely, the single aspects of the data (e.g. geometry, structure and so on) can be validated separately, possibly with automatic validators[15] whether they are quantitative or formalized.

4. A case study to exemplify the proposed approach.

A case study was chosen to exemplify and iteratively design the proposed approach in this paper. The goal of the integration is the update of a 3D city model (CityGML LoD0 and LoD1) (Figure 8) with the data coming from a BIM (IFC) representing the architectural model of a designed building, so called ‘Terraced tower’ (Figure 9), likely delivered for digital building permitting procedure. It is a rather simple case, to show how the proposed framework can be used for data integration in practice.

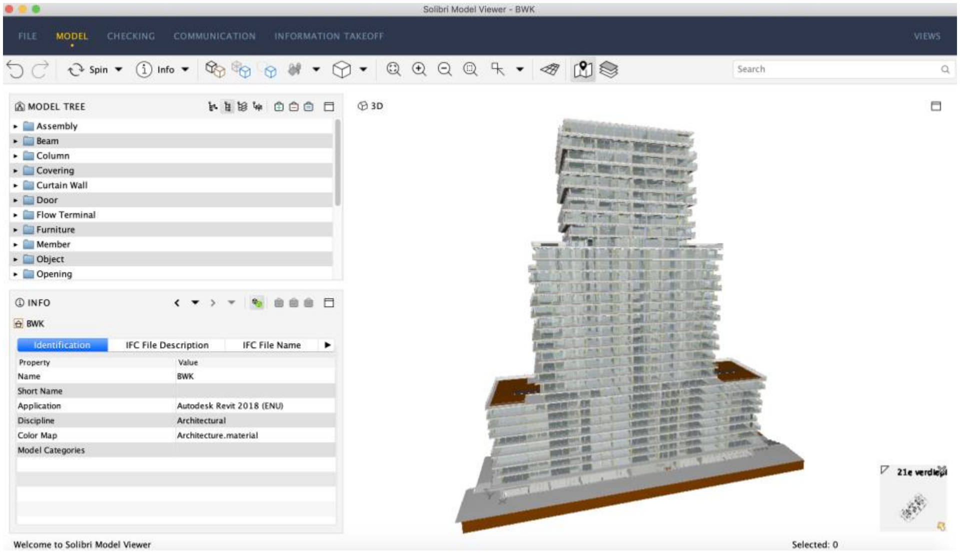
[15] E.g. http://geovalidation.bk.tudelft.nl [Accessed 1st December 2021]



- ☑ Building (bda9430e1-fe7d-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76a5add4-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76aa684f-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76ac15fc-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76acb2d2-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76ad278e-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76ad4ed0-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b76b05c6f-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b7faaca23-ff1a-11e5-8acc-1fc21a78c5fd-0)
- ☑ Building (b7faaca23-ff1a-11e5-8acc-1fc21a78c5fd-1)
- ☑ Building (b7fab6718-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b7fac2a3d-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b7fadff24-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b7faec247-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b7faee98a-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b7fb0e581-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fc282ed-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fc3e230-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fc605c3-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fc7176d-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fc7daaf-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fc8eb42-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fcb3521-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fcbf954-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fccbca8-ff1a-11e5-8acc-1fc21a78c5fd)
  - ☑ lod0FootPrint
  - ☑ lod0RoofEdge
  - ☑ lod1Solid
- ☑ Building (b8fccbcb5-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fcce2f2-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fce1c01-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fce90bd-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fd03e74-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fd0da74-ff1a-11e5-8acc-1fc21a78c5fd)
- ☑ Building (b8fd101a0-ff1a-11e5-8acc-1fc21a78c5fd)

Attribute	Value
measuredHeight	37.33

Figure 8. LoD1 CityGML model of Rotterdam visualized in azul.[16] The building in yellow will be updated and substituted after the integration.



MODEL TREE

- Assembly
- Beam
- Column
- Covering
- Curtain Wall
- Door
- Flow Terminal
- Furniture
- Member
- Object
- Opening

INFO

BWK

Identification | IFC File Description | IFC File Name

Property	Value
Name	BWK
Short Name	
Application	Autodesk Revit 2018 (ENU)
Discipline	Architectural
Color Map	Architecture.material
Model Categories	

Welcome to Solibri Model Viewer

Selected: 0

Figure 9. Architectural IFC model of the Terraced tower building, visualized in the Solibri Model Viewer (https://www.solibri.com/solibri-anywhere [Accessed 2nd December 2021]).

[16] https://github.com/tudelft3d/azul [Accessed 2nd December 2021].

4.1 Data requirements definition and integrability assessment

The data requirements, in this case, correspond with the characteristics of the destination model, i.e., the LoD1 CityGML model of Rotterdam. In Tables A1-A5, in Annex 1, each parameter is analysed for the destination model (data requirement) and in the input model, i.e., the IFC model of the Terrace tower building. In the last column, the integrability (and, consequently, needed harmonization processing) is assessed and commented according to the framework proposed in Section 3.5.

None of the scores given is 0, meaning that the two data sets can be integrated. Moreover, the minimum score is 2, so that the necessary processing is generalisation for some parameters, while others can be used as is or only converted to the destination format.

4.2 Processing of the IFC model towards harmonization

As pointed out by the specific assessment (Table A1), the geometry needs to be generalized, converted into meters from millimetres and stored into a different format.

For doing this, the IFC geometry was processed to extract the footprint of the building and the maximum height. Extract Transform and Load (ETL) tools can be used. In this case, the GEOBIM_Tool[17], developed for a project on the digitalization of the building permitting procedure in Rotterdam[18] (Noardo et al., 2022), was used to extract the footprint from the IFC model and to measure the building maximum height. The tool automatically scales the model into metre unit of measure.

[17] https://github.com/twut/GEOBIM_Tool [Accessed 2nd December 2021]

[18] https://3d.bk.tudelft.nl/projects/rotterdamgeobim_bp/ [Accessed 2nd December 2021]

The footprint was used to generate: the footprint polygon (lod0FootPrint) at the ground level; the roof edge polygon (lod0RoofEdge), generated by storing the same polygon at the height of the maximum height measured from the IFC model; an extruded solid representing the 3D building (lod1Solid). This processing embody the conversion step required.

The extrusion can be generated in any GIS or ETL processing tool from the footprint and maximum height of the building, using a similar approach as the one used for modelling the 3Dfier Rotterdam model. Finally, the geometries can be exported to GML.

Due to the chosen approach for the processing, the IFC semantics can be useful to select the objects that need to be considered in the processing. The objects which are not part of the *ifcBuilding*, such as the parts of the model belonging to the site, outside the building, need to be excluded from the geometric processing. The conversion necessary, in this case, must consider the entire representation paradigm (both geometrical and semantic/structural): we need to be aware that in the IFC model, *ifcBuilding* is a class that groups other building elements represented as objects with their own geometry, while in the CityGML model, it stores one only object with its own geometry(es).

In the storage of attributes and attribute values, there is no need for conversion, since, as resulted from the mapping, there is no attribute in IFC to store the maximum height of the building. However, the parameter measured from the IFC geometries is stored in the correct place to match the destination data encoding and syntax.

4.3 Final steps: data fusion, validation and metadata update.

The data need to be merged into each other. Assessing and resolving the conflicts is quite simple in this case. The data coming from the processing of the IFC model will substitute the building in place in the outdated 3D city model, to generate an up-to-date version of it. The bordering objects (e.g. roads, land cover) need to be modified to obtain a watertight model (Figure 10).

Versioning techniques need to be considered for keeping track of the integration as a change in the data.

Since the rsulting dataset is CityGML compliant, it can be validated with the online validator val3dity[19].

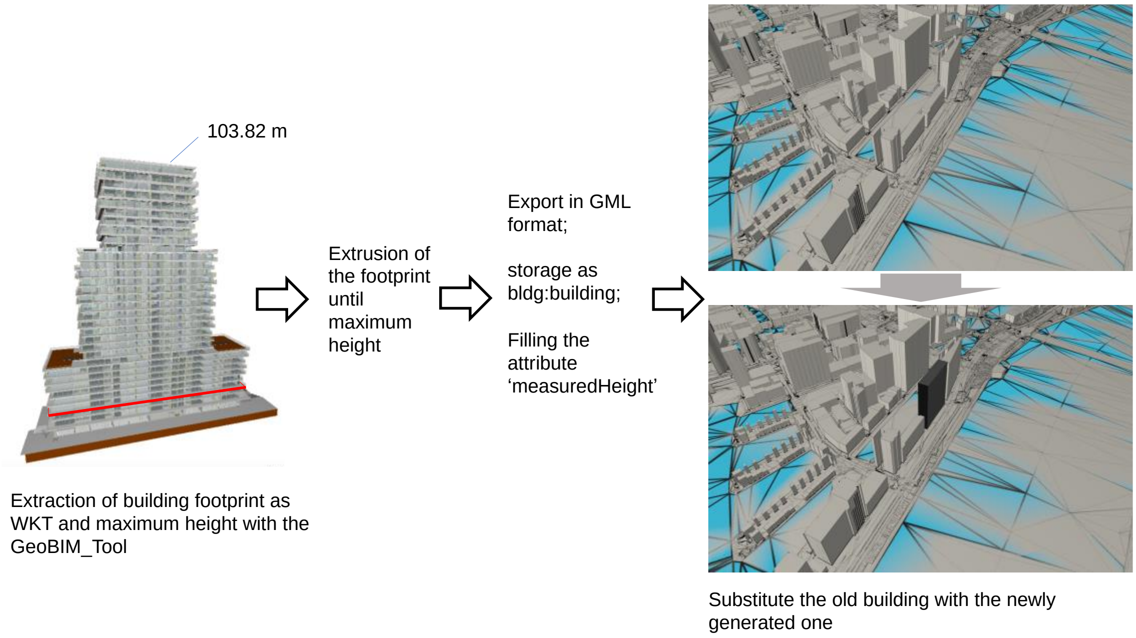

103.82 m
Extraction of building footprint as WKT and maximum height with the GeoBIM_Tool
Extrusion of the footprint until maximum height
Export in GML format;
storage as bldg:building;
Filling the attribute 'measuredHeight'
Substitute the old building with the newly generated one

Figure 10. Processing followed for the integration (harmonization + data merging) as planned according to the initial assessment.

A similar approach could be used also in more simple cases, for example, to update 2D digital maps (e.g. Figure 11)

[19] https://github.com/tudelft3d/val3dity Accessed 06/06/2022

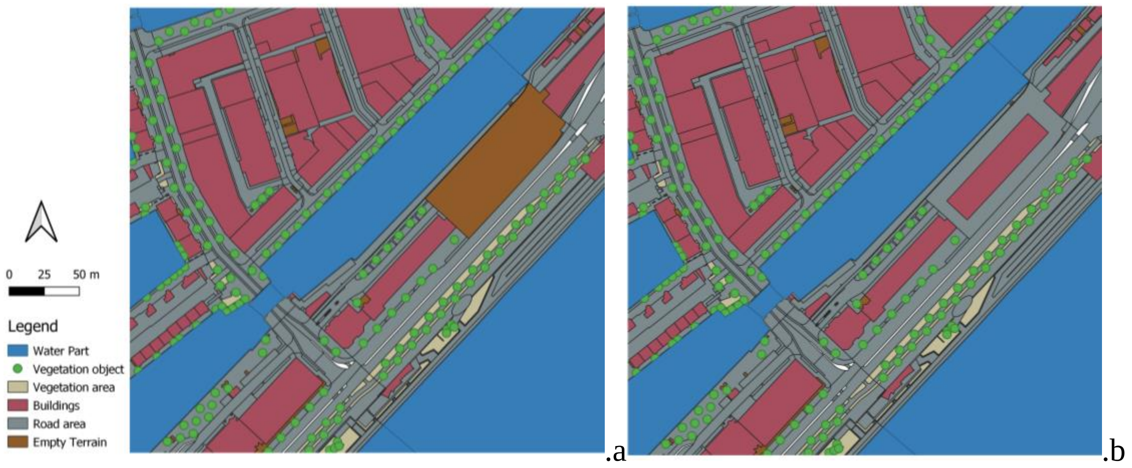


Figure 11. Example of the use of extracted data to update a 2D digital map (a. before; b. after the update), following the same workflow.

5. Discussion

The proposed methodology supports a point-by-point analysis to obtain an actual integration of datasets with respect to use case data requirements. The proposed framework joins the efforts made within different fields, such as ontology engineering, 3D survey and 3D city modelling.

Due to the high complexity of the issue, it is not possible to provide one only solution, but the overall framework and methodology is proposed, which was not available in literature before as a comprehensive workflow and reference, but with specific focus on single aspects. Some options for managing and processing the different aspects are suggested for each case, as proposed in literature. The needed and available options to process the data with respect to each outlined parameter must be investigated into detail for each case, since the range of existing data is too heterogeneous and the use case requirements may be very specific for each case, which makes any suggestion for specific processing not effective. However, the proposed framework can be used for any kind of integration, representing a solid reference for future work.

5.1 The automation of the methodology

The integration framework, as defined in this paper, is very complex, including qualitative assessments in some cases. Therefore, it is hard to propose a fully automatic procedure to integrate datasets. At this moment, the main aim of this approach is to have a transparent and clear methodology for effective data integration that can be reused in the future and at the same time highlight possible required improvements of individual data sources.

An automated workflow should be constituted by many components, in order to process the data with respect to each parameter, to be previously measured and assessed. The level of possible automation would increase according to metadata's quality (they should be correct and specific), completeness (all the parameters should be well documented and explained) and formality of the storage language (they should be machine-readable).

This would allow an algorithm to choose or suggest harmonization processing for each parameter and final data merging. However, at the moment, considering the available data (and metadata) from practice, automatic procedures can only be chosen to support the pre-processing harmonization steps for the single parameters. Moreover, some automatic or computer-assisted procedure can be used to extract the needed parameters from the data, whether these are not properly documented (e.g. Kavouras and Kokla, 2002). A manual or semi-manual guided analysis is still the preferable choice.

5.2 An overall workflow

The group of actions needed to convert the data into the integrated dataset can be implemented in different workflows. The kind of implementation and tools have to be chosen case by case, according to the complexity of the conversion, the level of possible automation and the need for repeating the process several times. They range from manual workflows, in which the actions intended to solve each of the detected inhomogeneities are launched manually step by step, to completely automated workflows, implemented in ETL tools, such as the Safe Software FME[20], in software like GIS processing models, or in other bespoke implementations.

Currently, many studies developing converters of data, for example between BIM and GIS already consider several of the listed aspects together, most frequently the data format and data schema/semantics are taken into account (Stouffs et al., 2018; Salheb et al., 2020; Clemen et al., 2021; Barbato et al., 2018) and in the most advanced cases also the transformation of geometries according to the required representation paradigm (Donkers et al., 2016; Olsson et al., 2019).

Few examples are available about the complete workflow from data assessment to pre-processing and harmonization and final data merging into real integration.

5.3 Discussion about metadata

In order to allow a fast assessment, suitable metadata should contain up-to-date and specific information about all the mentioned points. However, many of them are not foreseen in the current standard metadata schemas.

[20] https://www.safe.com [Accessed 3rd December 2021]

Table 7 maps the metadata of some of the most popular Open standard data model to the parameters defined in Section 3.1. ISO 19115[21] is the reference standard about metadata for geographic information. Most of other data models are compliant to it (INSPIRE, CityJSON), but they can present some variations if they include additional attributes or exclude some of the ISO attributes from the profile they use.

LandInfra standard[22] is ISO19115 compliant as well (Kumar et al., 2019).

CityGML is not included in Table 7 because it has limited metadata support (i.e., only name of the dataset, bounding box, coordinate system), mostly optional and inherited from the GML encoding format. In some cases, more information is added in form of comments in the XML file, but there is no control and no shared prescription about what information must be provided and in what format (Labetski et al., 2018). Metadata support was not included in the version 3 of CityGML either.

Similarly, many IFC files can contain some additional metadata information in a commented section in the STEP file in which they are encoded (e.g., author, schema, creation date).

The HEADER of the STEP files can also host some additional information (e.g., description, time_stamp, author, originating_system, FILE_SCHEMA). However, they do not follow a specific prescription and are related to the generation of the STEP file, according to the specific implementation choices of software exporting it.

Table 7 – Mapping of the defined parameters to Open standards' metadata schemas. In green cells, fully compliant information with the parameters considered in this paper can be found. In orange cells, some related information to the parameters is foreseen, or the link to unstructured documents within which more attributes can be searched.

[21] https://www.iso.org/standard/53798.html [Accessed 24th November 2021]

[22] https://www.ogc.org/standards/landinfra [Accessed 24th November 2021]

		ISO19115	INSPIRE[23]	IFC	CityJSON
Contents and procedures	*Spatial extent*	Geographic location of the dataset	geographic bounding box		geographicalExtent
	Temporal frame	Dataset reference date, Additional extent information for the dataset	temporal reference, temporal extent, date of last revision, date of creation		datasetReferenceDate
	Scope	Dataset topic category	resource type, topic category	IDM, IDS	datasetTopicCategory
	Goal and use case requirements	Abstract describing the dataset	Resource abstract	IDM, IDS	Abstract, specificUsage
	Lineage	Lineage	Lineage		lineage
	Author	Dataset responsible party	Responsible organization		
	Implementation requirements		specifications[24]		spatialRepresentationType
Geometry	*Accuracy*		specifications		
	Abstraction level	Spatial resolution of the dataset	Spatial resolution		presentLoDs
	Geometry paradigm	Spatial representation type	specifications	IDM, IDS, MVD	

[23] The INSPIRE metadata model is based on ISO19115, ISO19119 and ISO 15836 (Dublin Core) (ECJRC, 2010)

[24] Reference to an external document where all detailed specifications can be described. However, there is no guideline about what to include in such specifications.

	Topology		specifications	IDM, IDS, MVD	
	Georeferencing	Reference system	specifications	IfcMapConversion	referenceSystem
	Unit of measure		specifications	IfcProject - ProjectUnits	
Semantics	*Entities*		specifications	IDM, IDS, MVD	
	Properties and attributes		specifications	IDM, IDS, MVD	
	Codelists and values		specifications	IDM, IDS, MVD	
	Terms		specifications	IDM, IDS, MVD	
	Accuracy (vagueness)		specifications		
	Approximation level		specifications		
	Semantic paradigm		specifications	IDM, IDS, MVD	
	Language	Dataset language	resource language		datasetLanguage
	Encoding		specifications		
Structure	*Is-a hierarchies;*		specifications	IDM, IDS, MVD	
	Part-of meronymic hierarchies;		specifications	IDM, IDS, MVD	
	Relationships		specifications	IDM, IDS, MVD	
	Reference data model,		specifications	IDM, IDS, MVD	
	version		specifications	IDM, IDS, MVD	
	profile		specifications	IDM, IDS, MVD	
	extensions		specifications	IDM, IDS, MVD	
	Granularity		specifications	IDM, IDS, MVD	
Syntax	*Data format*		specifications	IDM, IDS, MVD	
	Objects' behavior		specifications	IDM, IDS, MVD	

In Table 7, only the metadata regarding the technical details of the data are considered, while others, related to non-technical aspects (licencing, use and retrieval of data, publication details) are not reported.

As already pointed out by Labetski et al. (2018), many attributes about the datasets, which could be relevant for their correct interpretation and (re)use, as well as for their integration, are currently missing in the official metadata schemas. They link to external specifications document, in some cases (e.g. INSPIRE) but without guaranteeing or prescribing anything about the information there contained.

In the table, we can notice how only a minimal part of the useful attributes is covered by metadata schemas. Both INSPIRE and IFC foresee the use of external documents to specify further information. The buildingSMART standards related to IFC propose a formal way to encode such information, while no guidelines are prescribed by the INSPIRE data model. However, there is currently little specification about what information should be stored in the documents or files and how, since many buildingSMART standards are still in the process of being defined. In addition, there are few examples of the use of such tools in practice. The use of metadata in general, in practice, is often still neglected.

6. Conclusions

The topic of multisource spatial data integration is relevant to many use cases applications, such as GeoBIM, digital twins, governance digitalization, land analysis. GeoBIM is the specific integration of geoinformation with BIM and can, in turn, serve several use cases (digital building permits, maps updating, asset and facility management and so on). Digital twins need to integrate several kinds of data within the same system and, according to the use case chosen, such data must comply to specific requirements and likely be integrated within the same dataset for analysis purposes. Many other kinds of land analysis need the integration of various data, which are likely distributed across several datasets. Some examples are: climate and microclimate analysis (3D geoinformation, terrain, soil model, whether data, vegetation, buildings materials in the most advanced cases); noise analysis (3D geoinformation, terrain, functions, traffic data, noise barriers parameters etc.); road infrastructure maintenance analysis (geoinformation, traffic data, transport infrastructure details). Data integration allows saving resources to generate new data by re-using the existing data sets, as well as to enable automation of several tasks. However, the high level of complexity of 3D information systems, such as BIM or 3D city models, and their, even conceptual, distance to each other make the integration workflows hard. As a consequence, the integration attempts often remain partial. This paper provides a reference for spatial data integration to support use cases applications, by proposing a comprehensive workflow and methodology considering in detail all the data parameters involved in the integration: geometry, semantics, structure and syntax. By following the provided methodology, a proper and consistent harmonization and merging processing can be planned, to obtain integrated datasets usable in practice.

Because of the complexity of the involved components, this study could not provide one final solution for each of the parameters and the workflow steps. Moreover, the software tools to process the data are continuously evolving, as well as the modes to edit the models. Therefore, mentioning specific solutions would have been reductive with respect to the range of possibilities available now and in the future.

This paper outlines a workflow and framework guiding a suitable multisource data integration by considering the needs of use cases applications as well as the usual characteristics of the data sets that can be found in practice. Following the described workflow will allow obtaining data concretely re-usable, by means of a systematic approach, without neglecting any of the features defining the data and avoiding therefore any issue at the moment of re-using them within applications.

Future work will be directed at testing in more detail each variable detected in the assessment matrixes and testing the methodology with more cases and more complex data.

References

Ahn, D., Claridades, A.R.C., and Lee, J., 2020. Integrating Image and Network-Based Topological Data through Spatial Data Fusion for Indoor Location-Based Services, *Journal of Sensors*, 2020. DOI:10.1155/2020/8877739

Arroyo Ohori K, Diakité A, Krijnen T, Ledoux H, and Stoter J., 2018. Processing BIM and GIS Models in Practice: Experiences and Recommendations from a GeoBIM Project in The Netherlands. *ISPRS International Journal of Geo-Information*. 7(8):311. DOI:10.3390/ijgi7080311

Arroyo Ohori, K., 2020. azul: A fast and efficient 3D city model viewer for macOS. *Transactions in GIS*, 24(5), 1165-1184. DOI: 10.1111/tgis.12673

Baig, S.U. and Abdul-Rahman, A., 2013a. Generalization of buildings within the framework of CityGML, *Geo-spatial Information Science*, 16(4), 247-255, DOI:10.1080/10095020.2013.866617

Baig, S. U., and Abdul-Rahman, A., 2013b. Generalization and visualization of 3D building models in CityGML. In:*Progress and New Trends in 3D Geoinformation Sciences*. 63-77. Springer, Berlin, Heidelberg.

Barbato, D., Pristeri, G., and De Marchi, M., 2018. GIS-BIM Interoperability for Regeneration of Transurban Areas. In: *REAL CORP 2018–EXPANDING CITIES–DIMINISHING SPACE. Are "Smart Cities" the solution or part of the problem of continuous urbanisation around the globe? Proceedings of 23rd International Conference on Urban Planning, Regional Development and*

Information. 243-250. CORP–Compentence Center of Urban and Regional Planning.

Biljecki, F., Heuvelink, G. B. M., Ledoux, H., and Stoter, J., 2015. Propagation of positional error in 3D GIS: Estimation of the solar irradiation of building roofs. *International Journal of Geographical Information Sciences*, 29(12), 2269–2294. DOI:10.1080/ 13658816.2015.1073292

Biljecki, F., Ledoux, H., and Stoter, J., 2016. An improved LOD specification for 3D building models. *Computers, Environment, and Urban Systems*, 59, 25-37.

Biljecki, F., and Dehbi, Y., 2019. Raise the roof: towards generating LoD2 models without aerial surveys using machine learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*.

Biljecki, F., and Tauscher, H., 2019. Quality of BIM–GIS conversion. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*.

Bloch, T., and Sacks, R., 2018. Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models. *Automation in Construction*, *91*, 256-272. DOI:10.1016/j.autcon.2018.03.018

Chen, H., Sun, Q., Xu, L., and Xiong, Z., 2013. Application of data fusion in the production and updating of spatial data. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *7*, W1.

Clemen, C. and Görne, H., 2019. Level of Georeferencing (LoGeoRef) using IFC for BIM. *Journal of Geodesy, Cartography and Cadastre. 10*, 15–20.

Clemen, C., Kaiser, T., Romanschek, E., and Schröder, M., 2021. Site Plan for BIM?–A Free and Open Source Plug-In for As-Is Vicinity Models to be Used in Small and Medium-Sized BIM-Projects. In:*FIG e-Working Week 2021 Smart*

Surveyors for Land and Water Management - Challenges in a New Reality Virtual, 21–25 June 2021.

Cockcroft, S., 1997. A Taxonomy of Spatial Data Integrity Constraints. *GeoInformatica* 1, 327–343. DOI:10.1023/A:1009754327059

Colucci, E., Kokla, M., and Norado, F., 2020. Semantic Comparison of 3D City Datasets and Mapping to Geospatial Ontologies. *Conference presentation, in 3DGeonfo Conference 2020*, Virtual Event, September 2020.

Dabove, P., and Di Pietra, V., 2019. Single-Baseline RTK Positioning Using Dual-Frequency GNSS Receivers Inside Smartphones. *Sensors*. 19(19):4302. DOI:10.3390/s19194302

Dextre Clarke, S., 2011. ISO 25964: a standard in support of KOS interoperability. In *Facets of knowledge organization: proceedings of the ISKO UK Second Biennial Conference, 4th-5th July 2011, London.* 129-134.

Dextre Clarke, S. G., and Zeng, M. L., 2012. From ISO 2788 to ISO 25964: The evolution of thesaurus standards towards interoperability and data modelling. *Information Standards Quarterly (ISQ)*, *24*(1).

Devogele, T., Parent, C., and Spaccapietra, S., 1998. On spatial database integration. *International Journal of Geographical Information Science*, *12*(4), 335-352. DOI: 10.1080/136588198241824

Devogele, T., 2002. A new merging process for data integration based on the discrete Fréchet distance. *Advances in spatial data handling*. Springer, Berlin, Heidelberg. 167-181.

Diakité, A.A., Damiand, G., van Maercke, D., 2014. Topological Reconstruction of Complex 3D Buildings and Automatic Extraction of Levels of Detail. In:

Eurographics Workshop on Urban Data Modelling and Visualisation, Strasbourg, France. 25-30, DOI:10.2312/udmv.20141074.

Doan, A., Madhavan, J., Domingos, P., and Halevy, A., 2004. Ontology matching: A machine learning approach. In: *Handbook on ontologies.* Springer, Berlin, Heidelberg. 385-403.

Doerr, M., 2001. Semantic problems of thesaurus mapping. *Journal of Digital information*, *1*(8), 2001-03.

Doerr, M., 2004. *Semantic interoperability: theoretical considerations*. TR 345, Institute of Computer Science, FORTH, Science and technology Park of Crete, Crete, Greece.

Dong, W., Wang, Q., Wang, X., and Zha, H., 2018. PSDF fusion: Probabilistic signed distance function for on-the-fly 3D data fusion and scene reconstruction. In: *Proceedings of the European Conference on Computer Vision (ECCV)* 701-717.

Donkers, S., Ledoux, H., Zhao, J., and Stoter, J., 2016. Automatic conversion of IFC datasets to geometrically and semantically correct CityGML LOD3 buildings. *Transactions in GIS*, *20*(4), 547-569.

Dou, D., Wang, H., and Liu, H., 2015. Semantic data mining: A survey of ontology-based approaches. In: *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 2015, 244-251, DOI: 10.1109/ICOSC.2015.7050814.

Duchêne, C., Baella, B., Brewer, C. A., Burghardt, D., Buttenfield, B. P., Gaffuri, J., Käuferle, D., Lecordix, F., Maugeais, E., Nijhuis, R., Pla, M., Post M., Regnauld, N., Stanislawski, L. V., Stoter, J., Tóth, K., Urbanke, S., van Altena, V., and Wiedemann, A., 2014. Generalisation in practice within national mapping agencies. In *Abstracting geographic information in a data rich world*. 329-391. Springer, Cham.

Duckham, M., and Worboys, M., 2005. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science*, 19(5), 537-557. DOI: 10.1080/13658810500032339

ECJRC - Drafting Team Metadata and European Commission Joint Research Centre, 2010. *INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119.* INSPIRE MD_IR_and_ISO_v1_2_20100616

Egenhofer, M. J., and Frank, A., 1992. Object-oriented modeling for GIS. *Journal of the Urban and Regional Information Systems Association*, *4*(2), 3-19.

Egenhofer, M.J., Clementini, E., and di Felice, P., 1994. Research Paper, International Journal of Geographical Information Science, 8:2, 129-142, DOI: 10.1080/02693799408901990

Ellul, C., Noardo, F., Harrie, L., and Stoter, J., 2020. the EuroSDR GeoBIM project – developing case studies for the use of GeoBIM in practice, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIV-4/W1-2020, 33–40, DOI:10.5194/isprs-archives-XLIV-4-W1-2020-33-2020.

European Union, 2017. *New European Interoperability Framework – Promoting seamless services and data flows for European public administrations.* Luxembourg, Publication Office of the European Union. ISBN 978-92-79-63756-8 doi:10.2799/78681. Available at https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf Accessed 06/06/2022

Gaffuri, J., 2011. Improving web mapping with generalization. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *46*(2), 83-91.

Geiger, A., Benner, J., and Haefele, K. H., 2015. Generalization of 3D IFC building models. In: *3D geoinformation science*. 19-35. Springer, Cham.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., and Benediktsson, J.A., 2019. Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1), 6-39. DOI: 10.1109/MGRS.2018.2890023.

Girres, J. F., and Touya, G., 2010. Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, *14*(4), 435-459. DOI: 10.1111/j.1467-9671.2010.01203.x

Gröger, G., and Plümer, L., 2012. CityGML–Interoperable semantic 3D city models. *ISPRS Journal of Photogrammetry and Remote Sensing*, *71*, 12-33.

Guercke, R., and Brenner, C., 200. A Framework for the Generalization of 3D City Models. In: *Proceedings of 12th AGILE Conference on GIScience*, 31.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design*, *37*(4), 682-703.

Hiebel, G., Doerr, M. and Eide, Ø., 2017. CRMgeo: A spatiotemporal extension of CIDOC-CRM. *Int J Digit Libr* 18, 271–279. DOI:10.1007/s00799-016-0192-4

Jaud, Š., Donaubauer, A., Heunecke, O., and Borrmann, A., 2020 Georeferencing in the context of building information modelling. *Autom. Constr*. 118, 103211.

Jun, X., 2019. An Improved Spatial Topological Relationship Model and Algorithm. In *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)* IEEE. 993-997.

Kavouras, M., and Kokla, M., 2002. A method for the formalization and integration of geographical categorizations. *International Journal of Geographical Information Science*, *16*(5), 439-453.

Kavouras, M., and Kokla, M., 2007. *Theories of geographic concepts: ontological approaches to semantic integration*. CRC Press. ISBN: 978-0-8493-3089-6

Klein, M., 2001. Combining and relating ontologies: an analysis of problems and solutions. In: *Proc. IJCAI-01 Workshop: Ontologies and Information Sharing,* Seattle, WA.

Krijnen, T., Noardo, F., Ohori, K. A., Ledoux, H., and Stoter, J., 2020. Validation and Inference of Geometrical Relationships in IFC. In *Proceedings of the 37th International Conference of CIB W*. 78.

Kumar, K., Labetski, A., Ohori, K., Ledoux, H., and Stoter, J., 2019. The LandInfra standard and its role in solving the BIM-GIS quagmire. *Open geospatial data, softw. stand.* 4(5). DOI:10.1186/s40965-019-0065-z

Labetski, A., Kumar, K., Ledoux, H., and Stoter, J., 2018. A metadata ADE for CityGML. *Open geospatial data, softw. stand.* 3(16). DOI:10.1186/s40965-018-0057-4

Latiffi, A. A., Brahim, J., Mohd, S., and Fathi, M. S., 2015. Building information modeling (BIM): exploring level of development (LOD) in construction projects. In *Applied Mechanics and Materials.* 773. Trans Tech Publications Ltd. 933-937.

Laurini, R., and Thompson, D., 1992. *Fundamentals of spatial information systems.* Academic press. 37. ISBN: 0-12-438380-7

Laurini, R., 1998. Spatial multi-database topological continuity and indexing: a step towards seamless GIS data interoperability. *International Journal of Geographical Information Science,* 12(4), 373-402, DOI:10.1080/136588198241842

Ledoux, H., Arroyo Ohori, K., 2017. Solving the horizontal conflation problem with a constrained Delaunay triangulation. J Geogr Syst 19, 21–42. DOI:10.1007/s10109-016-0237-7

Ledoux, H., 2018. val3dity: validation of 3D GIS primitives according to the international standards. *Open geospatial data, softw. stand.* 3(1). DOI:10.1186/s40965-018-0043-x

Ledoux, H., Arroyo Ohori, K., Kumar, K., Dukai, B., Labetski, A., and Vitalis, S., 2019. CityJSON: A compact and easy-to-use encoding of the CityGML data model. *Open Geospatial Data, Software and Standards*, *4*(1), 1-12.

Ledoux, H., Biljecki, F., Dukai, B., Kumar, K., Peters, R., Stoter, J., and Commandeur, T., 2021. 3dfier: automatic reconstruction of 3D city models. *Journal of Open Source Software*, *6*(57), 2866.

Ledoux, H., Biljecki, F., Dukai, B., Kumar, K., Peters, R., Stoter, J., and Commandeur, T., 2021. 3dfier: automatic reconstruction of 3D city models. *Journal of Open Source Software*, *6*(57), 2866.

Lenzerini, M., 2002. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* 233-246. DOI:10.1145/543613.543644

Lunetta, R., Congalton, R., Fenstermaker, L., Jensen, J., Mcgwire, K., and Tinney, L. R., 1991. Remote sensing and geographic information system data integration-Error sources and research issues. *Photogrammetric engineering and remote sensing*, *57*(6), 677-687.

Lüscher, P., Burghardt, D., Weibel, R., 2007. *Ontology-driven enrichment of spatial databases*. In: *10th ICA Workshop on Generalisation and Multiple Representation*, Moskau, 2 August 2007 - 3 August 2007. International Cartographic Association, online.

Malinowski, E., and Zimányi, E., 2006. Requirements specification and conceptual modeling for spatial data warehouses. In: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* Springer, Berlin, Heidelberg. 1616-1625.

McGuinness, D. L., Fikes, R., Rice, J., and Wilder, S., 2000. An environment for merging and testing large ontologies. In: Cohn, A.G., Giunchiglia, F., and Selman, B., editors, *KR2000: Principles of Knowledge Representation and Reasoning*, 483–493, San Francisco, 2000. Morgan Kaufmann.

Mohammadi, H. Binns, A. Rajabifard, A., and Williamson, I. P., 2006. Spatial Data Integration. In: *Seventeenth UN Regional Cartographic Conference for Asia and the Pacific*, Bangkok. http://hdl.handle.net/11343/26704 [Accessed 2nd December 2021]

Mohammadi, H., Rajabifard, A., and Williamson, I.P., 2010. Development of an interoperable tool to facilitate spatial data integration in the context of SDI, *International Journal of Geographical Information Science*, 24(4), 487-505, DOI: 10.1080/13658810902881903

Morocho, V., Saltor, F., and Pérez-Vidal, L., 2003. Schema Integration on Federated Spatial DB Across Ontologies. In Engineering Federated Information Systems, In: *Proceedings of the 5th Workshop EFIS* 2003 ISBN 1-58603-359-X. 63-72.

Mostafavi, M. A., 2006. Semantic similarity assessment in support of spatial data integration. In: *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*.

Nguyen, H., Cressie, N., and Braverman, A., 2012. Spatial Statistical Data Fusion for Remote Sensing Applications. *Journal of the American Statistical Association*, 107:499, 1004-1018, DOI: 10.1080/01621459.2012.694717

Noardo, F., Lingua, A., Aicardi, I., and Vigna, B., 2016. Cartographic data harmonisation for a cross-border project development. *Applied Geomatics*, *8*(3), 133-150. DOI:10.1007/s12518-016-0172-9

Noardo, F., Ellul, C., Harrie, L., Overland, I., Shariat, M., Arroyo Ohori, K., and Stoter, J., 2020a. Opportunities and challenges for GeoBIM in Europe: developing a building permits use-case to raise awareness and examine technical interoperability challenges. *Journal of Spatial Science*, 65(20, 209-233, DOI: 10.1080/14498596.2019.1627253

Noardo, F., Harrie, L., Arroyo Ohori, K., Biljecki, F., Ellul, C., Krijnen, T., Eriksson, H., Guler, D., Hintz, D., Jadidi, M.A., Pla, M., Sanchez, S., Soini. V.-P., Stouffs, R., Tekavec, J., and Stoter, J., 2020b. Tools for BIM-GIS Integration (IFC Georeferencing and Conversions): Results from the GeoBIM Benchmark 2019. *ISPRS International Journal of Geo-Information*. 9(9):502. DOI:10.3390/ijgi9090502

Noardo, F., Krijnen, T., Arroyo Ohori, K., Biljecki, F., Ellul, C., Harrie, L., Eriksson, H., Polia, L., Salheb, N., Tauscher, H., van Liempt, J., Goerne, H., Hintz, D., Kaiser, T., Leoni, C., Warchol, A., and Stoter, J., 2021a. Reference study of IFC software support: the GeoBIM benchmark 2019 – Part I. *Transactions in GIS* 25(2) 805–841. DOI:10.1111/tgis.12709

Noardo, F., Arroyo Ohori, K., Biljecki, F., Ellul, C., Harrie, L., Krijnen, T., Eriksson, H., van Liempt, J., Pla, M., Ruiz, A., Hintz, D., Krueger, N., Leoni, C., Leoz, L., Moraru, D., Vitalis, S., Willkomm, P., and Stoter, J., 2021b. Reference study of CityGML software support: the GeoBIM benchmark 2019 – Part II.*Transactions in GIS* 25(2), 842–868. DOI:10.1111/tgis.12710

Noardo F, Arroyo Ohori K, Krijnen T, and Stoter J. 2021c. An Inspection of IFC Models from Practice. *Applied Sciences*. 11(5):2232. DOI:10.3390/app11052232

Noardo, F., Wu, T., Arroyo Ohori, K., Krijnen, T., and Stoter, J., 2022. IFC models for semi-automating common planning checks for building permit. *Automation in Construction.* 134, 104097. DOI: 10.1016/j.autcon.2021.104097

Noskov, A., and Doytsher, Y., 2017. A Linear Approach to Improving the Accuracy of City Planning and OpenStreetMap Road Datasets. *International Journal on Advances in Systems and Measurements,* 10(1and2), 23–34. DOI:10.5281/zenodo.1314636

OGC – Open Geospatial Consortium, 2012. *OGC City Geography Markup Language (CityGML) Encoding Standard.* OGC 12-019 Version 2.0.0. http://www.opengis.net/spec/citygml/2.0 [Accessed 2nd December 2021]

Olsson, P. O., Johansson, T., Eriksson, H., Lithen, T., Bengtsson, L. H., Axelsson, J., Roos, U., Neland, K., Rydén, B., Harrie, L., 2019. Unbroken digital data flow in the built environment process–a case study in Sweden. *International Archives of* *the Photogrammetry, Remote Sensing and Spatial Information Sciences*. XLII-2/W13, 1347–1352. DOI:10.5194/isprs-archives-XLII-2-W13-1347-2019, 2019.

Osman, I., Yahia, S. B., and Diallo, G., 2021. Ontology integration: Approaches and challenging issues. *Information Fusion*, *71*, 38-63. DOI:10.1016/j.inffus.2021.01.007

Park, Y., and Guldmann, J. M., 2019. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, environment and urban systems*, *75*, 76-89. Doi: 10.1016/j.compenvurbsys.2019.01.004

Ramos, M. M., and Remondino, F., 2015. Data fusion in cultural heritage-A review. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *40*(5), 359.

Rodriguez, M. A., and Egenhofer, M. J., 2003. Determining semantic similarity among entity classes from different ontologies. In: *IEEE Transactions on Knowledge and Data Engineering*, 15(2), 442-456. DOI: 10.1109/TKDE.2003.1185844.

Perumal, M., Velumani, B., Sadhasivam, A., and Ramaswamy, K., 2015. Spatial Data Mining Approaches for GIS – A Brief Review. In: Satapathy S., Govardhan A., Raju K., Mandal J. (eds) *Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Advances in Intelligent Systems and Computing*, vol 338. Springer, Cham. DOI:10.1007/978-3-319-13731-5_63

Salheb, N., Arroyo Ohori, K., and Stoter, J., 2020. Automatic conversion of CityGML to IFC, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIV-4/W1-2020, 127–134. DOI:10.5194/isprs-archives-XLIV-4-W1-2020-127-2020, 2020.

Sheeren, D., Mustière, S., and Zucker, J. D., 2004. How to integrate heterogeneous spatial databases in a consistent way?. In *East European Conference on Advances in Databases and Information Systems* Springer, Berlin, Heidelberg. 364-378

Spyns, P., Meersman, R., and Jarrar, M., 2002. Data modelling versus ontology engineering. *ACM SIGMod Record*, *31*(4), 12-17. DOI: 10.1145/637411.637413

Stoter, J., Post, M., van Altena, V., Nijhuis, R., and Bruns, B., 2014. Fully automated generalization of a 1: 50k map from 1: 10k data. *Cartography and Geographic Information Science*, *41*(1), 1-13.

Stouffs, R., Tauscher, H., Biljecki, F., 2018. Achieving Complete and Near-Lossless Conversion from IFC to CityGML. *ISPRS International Journal of Geo-Information*. 7(9):355. DOI:10.3390/ijgi7090355

Sun, J., Olsson, P.O., Eriksson, H., and Harrie L., 2019. Evaluating the geometric aspects of integrating BIM data into city models, *Journal of Spatial Science*, DOI:10.1080/14498596.2019.1636722

Thapa, K., and Bossler, J., 1992. Accuracy of spatial data used information systems. *Photogrammetric Engineering and Remote Sensing*, 58(6), 835-841.

Worboys, M. F., and Duckham, M., 2004. *GIS: a computing perspective*. CRC press.

Uggla, G., Horemuz, M., 2018. Geographic capabilities and limitations of Industry Foundation Classes. *Autom. Constr*. 96, 554–566.

Uitermark, H.T., van Oosterom, P.J.M., Mars, N.J.I., Molenaar, M., 2005. Ontology-based integration of topographic data sets, *International Journal of Applied Earth Observation and Geoinformation*, 7(2), 97-106, ISSN 0303-2434, DOI:10.1016/j.jag.2005.03.002.

Ulubay, A., and Altan, M. O., 2002. A different approach to the spatial data integration. *International Archives of Photogrammetry Remote Sensing And Spatial Information Sciences*, *34*(4), 656-661.

Van den Brink, L., Stoter, J., and Zlatanova, S., 2013a. Establishing a national standard for 3D topographic data compliant to CityGML, International Journal of Geographical Information Science, 27:1, 92-113, DOI: 10.1080/13658816.2012.667105

Van den Brink, L., Stoter, J., and Zlatanova, S., 2013b. UML-based approach to developing a CityGML application domain extension. *Transactions in GIS*, *17*(6), 920-942. DOI:10.1111/tgis.12026

Van Heerden, N., 2021. *BIM and 3D City Models as Input for Microclimate Simulation*. MSc thesis in geomatics, Delft University of Technology. http://resolver.tudelft.nl/uuid:630d57be-5660-4971-84c2-83bf12b1d204 [Accessed 2nd December 2021]

Vitalis, S., Arroyo Ohori, K., and Stoter, J., 2019. Incorporating topological representation in 3D city models. *ISPRS International Journal of Geo-Information*, *8*(8), 347.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S., 2001. Ontology-based integration of information-a survey of existing approaches. In *Proc. IJCAI-01 Workshop: Ontologies and Information Sharing,* Seattle, WA, ed. H. Stuckenschmidt. 108-117.

Wallgrün, J. O., and Dylla, F., 2010. *A relation-based merging operator for qualitative spatial data integration and conflict resolution*. Technical Report 022-06/2010, Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition.

Wang, L., 2017. Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*, *3*(1), 8-15.

Werbrouck, J., Pauwels, P., Bonduel, M., Beetz, J., and Bekers, W., 2020. Scan-to-graph: Semantic enrichment of existing building geometry. *Automation in Construction*, *119*, 103286.

Wiemann, S., and Bernard, L., 2016 Spatial data fusion in Spatial Data Infrastructures using Linked Data. *International Journal of Geographical Information Science*, 30:4, 613-636, DOI: 10.1080/13658816.2015.1084420

Worboys, M. F., and Duckham, M., 2004. *GIS: a computing perspective*. CRC press. ISBN: 0-415-28375-2

Xue, F., Wu, L., and Lu, W., 2021. Semantic enrichment of building and city information models: A ten-year review. *Advanced Engineering Informatics*, *47*, 101245. DOI:10.1016/j.aei.2020.101245

Yin J, Fu P, Hamm NAS, Li Z, You N, He Y, Cheshmehzangi A, Dong J., 2021. Decision-Level and Feature-Level Integration of Remote Sensing and Geospatial Big Data for Urban Land Use Mapping. *Remote Sensing*. 13(8):1579. https://doi.org/10.3390/rs13081579

Zhang, J., 2010. Multi-source remote sensing data fusion: status and trends, *International Journal of Image and Data Fusion*, 1:1, 5-24, DOI: 10.1080/19479830903561035

Zhang B. et al. (2022) A Review of Data Fusion Techniques for Government Big Data. In: Liao X. et al. (eds) Big Data. BigData 2022. *Communications in Computer and Information Science*, vol 1496. Springer, Singapore. https://doi.org/10.1007/978-981-16-9709-8_4

Zhao L, Liu Z, Mbachu J., 2019. An Integrated BIM–GIS Method for Planning of Water Distribution System. *ISPRS International Journal of Geo-Information*. 8(8):331. DOI:10.3390/ijgi8080331

Zhu, Z., Donia, S., 2013. Spatial and visual data fusion for capturing, retrieval, and modeling of as-built building geometry and features. *Vis. in Eng.* 1(10). DOI:10.1186/2213-7459-1-10.

Annex 1 Explanation of the parameters considered for data harmonisation

The geometric parameters

Geometric accuracy measures the geometries positioning and resolution with respect to the ground reality (Girres, Touya, 2010; Worboys, Duckham, 2004; Laurini, Thompson, 1992).

The *level of abstraction* corresponds to the concept of cartographic generalization for traditional maps and can be seen as a joint concept of scale and resolution in traditional cartography (Laurini, Thompson, 1992). It allows representing the objects on the map applying the appropriate selection, simplification, symbolization and classification for them to be understandable for a specific scale and a specific purpose (Worboys, Duckham, 2004; Gaffuri, 2011; Stoter et al., 2014, Duchêne et al., 2014). In case of 3D models, the Level of Detail (LoD) concept applies, first defined by CityGML (Gröger, Plümer, 2012; Biljecki et al., 2016). Other kinds of simplification, such as the Level of Development used in Building Information Modelling (Latiffi et al., 2015), are not relevant in this context, since they do not represent an abstraction from a model most faithful to reality, but indicate instead the stage through the path of design and improvement. It should be therefore considered in the retrieval of data phase, to assess whether they are suitable to integration, but in case of BIM, the reference from which to abstract more generalised representations should usually be the final design or the as-built BIM.

The same *geometry* can be represented, modelled and stored following several alternatives or '*paradigms'* (raster, vector, implicit, explicit, boundary representation, solid, voxels, etc.). Different modelling options are typical for different types of data, for example, 3D city models usually adopt boundary representation explicit geometries, while BIM uses implicit parametrically modelled geometries (Composite Solid

Geometries, swept solids, NURBS, etc.) (Noardo et al., 2020, Arroyo Ohori et al., 2018, Ledoux, 2018). The applications using the geometry for analysis or further processing (i.e., not only for visualization) need specific input. Therefore, depending on the use cases, and as defined accordingly in the data requirements, the data to be integrated must use the same kind of representation and storage of the geometries, in order to be suitably recognised and used properly.

Topology can be within one object, as part of the storage and representation of geometries, and between two objects, as spatial relationships (Ledoux, 2018). These characteristics can have an influence on the use for which the models are intended, as well as other constraints (e.g., Cockcroft, 1997).

Consistent *georeferencing* is an extremely relevant premise of any integration. It must take into account the used coordinate reference systems, both for planar coordinates (X, Y) and for heights, including: datum, projection, coordinate system, accuracy, and measuring systems (precision, accuracy, reliability, etc.). For example, data acquired with smartphones’ GNSS sensors or from crowd mapping can have discrepancies with respect to similar data acquired by means of more precise instruments (Dabove, Di Pietra, 2019, Haklay, 2010) that can be relevant depending on use cases.

Unit of measure deals with making the represented objects homogeneous with respect to the scale or precision (Laurini, Thompson, 1992) with which they are represented. In some cases, on-the-fly transformations (e.g., in GIS) allow correct visualization and, more seldom, analysis of the data. However, in most of cases it would be necessary to re-scale the data to a same unit of measure.

The semantic parameters

Semantics consists of **entities**, or classes of represented objects, their attributes and the foreseen values of codelists, which are used in the models. Relationships between those are covered in the structural features.

Attributes are the thematic properties of objects. In turn, they can be represented by different terms and definitions, and can be filled by different values and according to various criteria. Types of values admitted and codelists used – with clear definitions and criteria or methods to use or calculate each value – must be clear. Explaining them with definitions and examples helps understanding the correct intended meaning and avoiding ambiguity.

Entities belonging to different datasets can be compared based on (in ascending analysis depth): the terms used; their definition; their properties and attributes; their instances; relationships.

Moreover, in the description and mapping of entities, attributes and codelist values, it is necessary to analyse the features listed in Table 1: used term, vagueness, approximation, semantic paradigm, language and encoding.

In some cases, geometric properties, such as spatial relationships and topology, could also be stored as semantics. It is relevant for integration to assess the mode of storage of these characteristics and consider it according to the data requirements definition. It could be in fact necessary to either remove some relationships explicitly stored as semantics (attributes or relationships), whether not necessary in the final data, or calculate and infer them and store them explicitly whether this need is foreseen. One example of this could be the grouping of storeys in IFC files, which could be essential for some applications, while not relevant for others (such as the case study considered in this paper).

'Term' is the name used to indicate each concept: entity (class), attributes, codelist values (Kavouras and Kokla, 2007), as defined within the ISO25964 (Dextre Clarke, Zeng, 2012, Dextre Clarke, 2011). *'Definitions'* help defining semantics in the least ambiguous way, and sometimes include examples which further clarify the meaning. These can be compared to each other to support concepts mapping and integration purposes (Kavouras and Kokla, 2007).

'Vagueness' (or *'semantic accuracy'*) is described by Kavouras and Kokla (2007) as 'the degree of inexactness, fuzziness or indeterminable character of geographic concepts, properties and relationships. Uncertainty, randomness and ignorance contribute to the parameter'. Vagueness refers to the inability to clearly understand the meaning of a concept in a context. Examples of meanings that might not always be clear are 'large', 'high', 'dense'. Storing materials in BIM as just 'wood' or 'glass' can be vague as well for construction-intended purposes.

A different kind of vagueness is related to the source of information. For example, data coming from inferences or enrichment processing of the data will be vaguer than data acquired by survey or authoritative sources.

'Ambiguity', in contrast, is related to the existence of more than one specific meaning, which can be interpreted in different ways. For example, an *ifcWall* can be either loadbearing or not. There are several ways to understand or specify this, for example, the specific attribute can be used, or it can be assessed based on the disciplinary model being considered (whether structural or architectural, for example). Other examples can be in the interpretation of aerial imagery, whether green roofs are represented, which can be interpreted as grass field, and similar cases. It can be solved by specifying the context. Context (Kavouras, Kokla, 2007) is the restricted conceptual milieu giving meaning to the concept expressed. In the data models, it is usually described in the definition of each term/entity/class. Constraining the interpretation of each description, likely with examples, is also important to obtain consistent data.

'Approximation' (or *'semantic abstraction level'*) has to do with the granularity of the conceptualization and the level of detail reached by the semantic description.

'Semantic paradigm' is the reference reality and perspective for the conceptualization of the semantic representation of the data (Klein, 2001; Kavouras and Kokla, 2007). Within this feature, we can also include the criteria used to fill the attribute values or methods to be used to calculate them, as well as the unit of measure used.

The structural parameters

The structure, or schema, of thematic information is described in the data model or the ontology followed by the data. Although being slightly different artifacts (Spyns et al., 2002), the principles and features on which the integration of data models or ontologies depends can be considered similar.

Ontology science (e.g., Kavouras, Kokla, 2007; Mostafavi, 2006) provides useful tools and theory with respect to data structure integration. In some cases, the foreseen situations for ontology integration are more complex than what is usually found in data models structuring data from practice (for example, multiple inheritance is hardly present, or impossible, in data from practice). However, the concepts formulated can be reused to guide the integration of different semantic structures.

Semantic '*relationships*' among two or more concepts include dependency-association (Kavouras, Kokla, 2007). The *'is-a hierarchies*' must be considered, including the distance of a node (a concept) from the root of the ontology or data model and possible multiple inheritance. Similarly, *'part-of meronymic hierarchies'* are relevant to assess similarity of two structures.

'Granularity', intended as the smallest and biggest objects represented is the parameter useful to compare different data structures, together with the *'paradigm*' according to which the reality is interpreted and conceptualized in the data schema. The two parameters must be considered for both the is-a hierarchies and the part-of meronymies.

The syntactical parameters

For the syntactic level, the '*format'* encoding the data is relevant (e.g., GML, JSON, STEP, TIFF, shapefile, ASCII and so on), including the version of the implementation languages used and all the conditions and choices possibly adopted.

National *language* used is relevant (English, Italian, Dutch, French…), as well as the *encoding* of the terms and values for entities, attributes and codelists (Klein, 2001). For example, the use of uppercase or lowercase letters and punctuation; storage of dates as dd/mm/yyyy or mm/dd/yyyy; codelists referring to a code (e.g. a number or an alphanumerical string) or containing the value directly, and so on.

'Objects' behaviour' (in object-oriented spatial databases, operations and axioms that define them) (Egenhofer, Frank, 1992) is also amongst the features relevant for integration, although models from practice hardly reach such complexity.

Annex 2 Detailed definition of data parameters and integrability assessment.

Table A1. Pre-assessment of integrability based on general data features.

	Destination data (data requirements)	**Input data**	**Assessment**

Geographical extent	<gml:Envelope srsDimension="3" srsName="urn:ogc:def:crs:EPSG::7415">
	<gml:lowerCorner>90000.000 434963.000 0</gml:lowerCorner>
	<gml:upperCorner>94000.000 437500.000 100</gml:upperCorner>
	</gml:Envelope>
		#6611258= IFCCARTESIANPOINT((93191049.7637025,436743261.,0.));
		#6611260= IFCDIRECTION((0.681487612701515,0.731829648029095,0.));
		#6611262= IFCAXIS2PLACEMENT3D(#6611258,#19,#6611260);
		#6611263= IFCLOCALPLACEMENT($,#6611262);
		#6611264= IFCSITE('0uc_lgsWD3QgU5TbtwvEPN',#41,'Default',$,'',#6611263,$,$,.ELEMENT.,(42,21,30,344238),(-71,-3,-35,-194702),5600.,$,$);
			They overlap (case 3 Figure 4)

Temporal frame	2017	2020	Input data are more recent and can be suitably used to update the map, which is the goal of the integration.
Scope	City	Building	It overlaps (case B Figure 4)

Goal and intended use	Visualization of city objects	Building design	-
Lineage	Automatically generated by 3Dfier, extruding the building footprints until their maximum height. Input data: Dutch national digital map Basisregistratie Addressen and Gebouwen (BAG)[25] and Actueel Hoogtebestand Nederland (AHN3)[26]	Parametrically modelled within BIM software	-
Implementation requirements	Will be visualised in CityGML viewers, such as Azul (Arroyo Ohori, 2020).	Needed to be managed in BIM software to support design and construction.	-
Author	3D geoinformation group – TU Delft	Provast – Architect: OZ Architects	-

Table A2. Integrability assessment based on the geometric parameters

	Destination data (data requirements)	**Input data**	**Assessment**
Accuracy	20 cm	1 cm (design precision is 1 mm, the value is lowered to take into account the possible discrepancy between design and construction)	4

[25] https://bag.basisregistraties.overheid.nl [Accessed 2nd December 2021]

[26] https://www.pdok.nl/introductie/-/article/actueel-hoogtebestand-nederland-ahn3- [Accessed 2nd December 2021]

Abstraction	lod0FootPrint; lod0RoofEdge; lod1Solid	Very high level of detail, interior, exterior, small elements and furniture are represented	2
Paradigm	lod0FootPrint – gml:MultiSurface	Parametrically modelled solids	3
Topology	no relevant info to be taken into account	-	-
Georeferencing	EPSG:7415	EPSG:7415	4
Units of measure	m	mm	3

Table A3. Integrability assessment based on the semantic parameters

	Destination data (data requirements)	**Input data**	**Assessment**
Entities			
Terms	WaterBody, Building, LandUse, PlantCover, Road – as defined by CityGML v.2	Many entities related to building and building elements (ifcPlate, ifcSlab, ifcWall, ifcStair, ifcWindow etc.) as defined by IFC v.2x3. “Building” is the term used for the entire building.	4
Vagueness	Definitions from CityGML v.2	Definitions from IFC documentation. The part of the model used is in this case the group of objects (IfcSpatialStructureElement)“IfcBuilding”[27]	4

[27] https://standards.buildingsmart.org/IFC/DEV/IFC4_3/RC1/HTML/schema/ifcproductextension/lexical/ifcbuilding.htm [Accessed 2nd December 2021]

Approximation	Building and city elements	Building elements, but a class exists for Building.	4
Semantic paradigm	According to the city representation scope	According to the building design and construction scope (compatible)	3
Attributes			
Terms	measuredHeight	Not stored explicitly	-
Vagueness	Very clear term	-	-
Approximation	Variations could be in the reference point measured (top of the roof, gutter level, intermediate point etc.)	-	-
Semantic paradigm	According to the city representation scope	-	-
Codelists	Not relevant/present	-	not relevant
Attribute values			
Terms	-	-	
Vagueness	We cannot know the accuracy of measurement, but the precision is 1 centimetre	We can measure maximum height with centimetre accuracy w.r.t. the designed building. However, this one should be measured and checked during the final construction testing and validation.	4
Approximation	-	-	-

Semantic paradigm	According to the city representation scope	According to the building design and construction scope (it implies that, for example, we could even store small details, such as roof furniture and installations)	3

Table A4. Integrability assessment based on the structural parameters

	Destination data (data requirements)	**Input data**	**Assessment**
Is-a hierarchy	Compliant to CityGML v.2	Compliant to IFC v.2x3	3
Granularity	Smallest object is the building	Smallest object is the building element	2
Semantic Paradigm	According to the city representation scope	According to the building design and construction scope	3
Part-of meronimy	Compliant to CityGML v.2	Compliant to IFC v.2x3	3
Granularity	Smallest object is the building	Smallest object is the space	2
Semantic Paradigm	According to the city representation scope	According to the building design and construction scope	3
Relationships	No relationships represented	-	-

Table A5. Integrability assessment based on the syntactical parameters

	Destination data (data requirements)	**Input data**	**Assessment**
Data format	GML3	STEP Physical File (IFC)	3
Language	English	English	4

Encoding	Entities: According to CityGML v.2 (bldg:Building) Attributes: Compliant to CityGML v.2 “measuredHeight” Values: Floating value Expressed in meters (decimal separator = point)	Entities: According to IFC v. 2x3 (ifcBuilding)	3
Objects’ behaviour	no objects’ behaviour present	-	-