Built environment data standards and their integration: an analysis of IFC, CityGML and LandInfra

Image credit: Bioregional

Version 1.1 • 28 January 2021
OGC Document 19-091r2
bSI TR1012

Contents

Authors

Thomas Gilbert, Newcastle University

Carsten Rönsdorf, Ordnance Survey

Jim Plume, buildingSMART Australasia

Scott Simmons, Open Geospatial Consortium

Nick Nisbet, AEC3

Hans-Christoph Gruler, Leica Geosystems

Thomas H. Kolbe, Technische Universität München

Léon van Berlo, buildingSMART International

For further information, please contact:

Open Geospatial Consortium: standards@opengeospatial.org

buildingSMART International: contact@buildingsmart.org

Summary

Demand for digital representations of built environments is accelerating and can only be satisfied through greater software interoperability and data integration. The objective of the Integrated Digital Built Environment (IDBE) joint working group is to address this challenge by bringing together experts from the Open Geospatial Consortium and buildingSMART to coordinate the development of the relevant data standards. This document is an output from IDBE in which we describe the state of three of the most prominent built environment standards – **CityGML**, **IFC** and **LandInfra** – and describe some of the problems that hinder their integration; finally, we propose actions points for overcoming these problems.

Key points

Predominant disparities that hinder integration:
- Intended general purpose, practical applications and types of objects intended to be modelled;
- Conceptualisation of real-world objects, their properties and their relationships;
- Formal languages used for conceptual modelling and the description of schemas;
- Spatial representation (geometric and geographic).

Fundamental operations that underpin use cases:
- The creation of and changes to digital representations of objects and their properties;
- Modelling of functional relationships between components of a digital representation;
- Visualisation of representations, including virtual and augmented reality;
- Querying of digital representations spatially and semantically.

Paradigms of integration in usage:
- All objects are represented digitally in (or converted to) instances of one schema;
- Instances of multiple schemas are federated temporarily into a single representation by a software environment – this representation may be spatial, semantic and/or temporal, involving renderings, tabular and/or graph representations;
- Elements contained within instances of multiple schemas are cross-referenced, providing links to additional information that is not explicitly defined in any one instance.

Proposed action points for overcoming integration challenges:
- Articulate in plain writing a set of illustrative use cases;
- Derive and make publicly available a shared vocabulary or definition dictionary;
- Author a best practice document on the use of three-dimensional georeferencing;
- Devise a system of common unique identifiers for real-world, physical objects;
- Agree on a collaborative mechanism for opportunistic harmonisation of conceptual representations.

1. Introduction and background

Integrated Digital Built Environment (IDBE) is a joint working group under buildingSMART International (bSI) and Open Geospatial Consortium (OGC) in which a vision for the future of digital built environments is being defined. IDBE's objective is to achieve better software interoperability and data integration in the geospatial and built environment domains through coordination of standards development activities.

This document is intended as a succinct and practical output from IDBE that i) contributes to an understanding of the existing state of the relevant data standards, ii) describes some of the problems that are hindering progress on their integration and iii) suggests an approach for overcoming these problems. The audience is both the working group itself and the broader built environment data modelling community. The content of this document comprises material from a breadth of contributors present at multiple collaborative events but primarily the IDBE workshop held in Munich in April 2018.

The geospatial and built environment domains have traditionally been regarded as distinct. The coarser resolution data that describe existing environments have been handled by GIS practitioners and their software; the finer resolution, prescriptive designs of future builds have been dealt with in the AEC industry by Building Information Modelling (BIM) specialists. The emergence of urban-scale GIS, retrospective BIM, the use of BIM for terrain and infrastructure modelling, and the near-ubiquity of real-time IoT analytics are examples of the progress that has blurred this domain boundary. The concept and realisation of digital twins depend on simultaneous representation of the indoor, outdoor, underground and over-ground environments within a single or federated model. Spatial data modelling is fundamental to much practice, research and policymaking. There is an increasing need for sharing and collaborative working and, due to the accelerating demand for digital representations of our built environment, this can only be satisfied through greater software interoperability and data integration.

Urban planning, management and operational activities require decision-making that accounts for information contained in more than one model. An architect might need to reposition windows in a BIM model in order to ensure a desirable view over a nearby green space that is only represented in an urban GIS model. A walkthrough simulation for the client would then need to render in 3D the geospatial context alongside the internal building layout for full appreciation of the design implications. The engineering contractor might require the integration of the external urban utility networks in order to identify suitable resource entry points to the building and ensure that cables and pipes of suitable material and dimensions are used; logical connections between the coarser urban and finer building scale networks would support the assessment of the building's suitability for electricity demand-side response and enable the detection of water leakage as it flows through the network. Post-construction, a building manager might need to query data describing operational efficiency and usage anomalies on a digital dashboard that is informed by an integrated model of the built environment.

The use cases for an integral digital built environment span a broad range of themes and applications. However, many of them relate to operations that are fundamentally very similar. Users need to be able to create and modify entities and their attributes/properties during planning, design and subsequent operation, and they need to be able to assert functional relationships between these components. Once the digital environment has been modelled, they need to be able to query its spatial and semantic content (subject to access rights) and conduct multi-dimensional and mixed reality visualisation, in order that value can be derived.

There are practical problems encountered when trying to model, simulate and analyse across multiple spatial scales, both before and after construction operations. It must be possible to integrate spatial information seamlessly and this depends on a level of commonality in the representation of the underpinning spatial data. This integration capability goes beyond mapping between data formats, discrepancies between local and geographic coordinate systems, and geometric levels of detail; it must consider the more fundamental interpretations of spaces and their relationships that underpin the conceptual models inherent to the formats.

Through the IDBE working group, we have identified three prominent, complementary and yet sometimes conflicting standards. These standards encompass many of the spatial scales and concepts necessary for modelling the built environment:

- **IFC** (Industry Foundation Classes) is used predominantly for exchange of rich, fine-scale building and infrastructure data in the AEC industry;
- **CityGML** (City Geography Markup Language) is a schema for interoperability and structuring urban data; and
- **LandInfra** (Land and Infrastructure Conceptual Model) has been introduced more recently to enable the modelling of land and civil engineering infrastructure facilities.

We describe the predominant disparities in purpose, scope, structure, overlaps and features of the three standards; we explain the challenges these disparities present to their integration; and, finally, we propose actions as a means of progressing.

2. Commonalities and differences

Although the CityGML, IFC and LandInfra standards have much in common, there are significant differences between them, which present challenges to integration. The nature of these differences and their origins need to be understood in order to appreciate and address the challenges faced.

Culture, practices, purpose and roles: the standards' purposes and the culture or practices of the industries in which they have gained prominence have influenced their divergence. While IFC was initially conceived for exchange of detailed building models, CityGML was devised to provide an analytical platform that allowed simulations within 3D city models, and LandInfra was introduced to address a capability gap in modelling land and engineering infrastructure facilities. CityGML is intended primarily to allow the description of urban environments as they can be observed or measured, whereas IFC and LandInfra are used primarily for design (or 'prescription') of how things will be. IFC is standardised by buildingSMART International (bSI) in collaboration with ISO, whereas CityGML is standardised by OGC and complies with its baseline - the complete set of specifications and standards that are approved by the OGC membership. LandInfra is also standardised by OGC and developed partly in collaboration with bSI. IFC serves primarily as an exchange format, with instances delivered throughout the lifecycle of a project and in parallel to 'working' versions in formats that are native to commercially available modelling software. It serves the multiple purposes demanded at different stages of design, construction and maintenance, supporting efficient change management provided by the BIM Collaboration Format (BCF). In contrast, CityGML is more often used as a native/working standard in which functions such as analysis, simulation, or visualisation are carried out, and spans a greater range of thematic domains. LandInfra was introduced as a use-case driven subset of and successor to the broadly scoped LandXML schema; a decision was made to develop a conceptual model and entirely new data standard, addressing a number of problems in LandXML such as the absence of a conceptual model, insufficient documentation, XML encoding issues and incompatibility with the OGC Baseline.

Thematic scope: this is the range of objects or types of environment that are representable. Both CityGML and LandInfra are well suited to the modelling of objects at relatively coarse spatial resolution. Through both its modules and extensions, CityGML provides substantial coverage at the urban scale of buildings, utility networks, energy and hydrology. LandInfra largely avoids overlap with CityGML and focuses more on the modelling of civil engineering works such as rail and road infrastructure, and concepts from the Land Administration Domain Model (LADM) such as cadastre and surveying. The predominant application of IFC is the modelling of finer-scale building and infrastructure components; and Mechanical, Electrical and Plumbing (MEP) components and networks in high geometric and semantic detail. Although there is less evidence of support for IFC at the urban scale, the standard has been extended to provide construction detail for roads, rails, bridges, tunnels, ports, waterways, landscape and urban design. These extensions complement the LandInfra standard, which provides for the modelling on these themes at a coarser resolution but very little on the building scale and nothing inside buildings. The thematic scopes of the three standards are depicted in Figure 1.

Figure 1 - Thematic coverage of the three data standards. Dark shading indicates strong coverage, light shading weaker coverage (or under development) and no shading implies no known coverage. The object icons are ordered approximately; the intention is to cluster by theme and represent finer spatial scales nearer the top.

Structure, sub-setting and extensibility: each of the standards comprises a core component plus additional components for common domains – these are termed modules (CityGML), parts (LandInfra) or domain-specific data schemas (IFC). CityGML can be extended by the user community for more specialist domains through the Application Domain Extension (ADE) mechanism; the InfraGML implementation of LandInfra can be extended similarly due to its XML base. IFC is extended through a formal buildingSMART process that involves user consensus and validation. It is possible for users to further extend the IFC core and its extensions using additional classifications and properties, and the buildingSMART Data Dictionary (bSDD) provides guidance on this. Where the entirety of the CityGML or LandInfra schemas are overly dense or complex, a subset may be chosen by specifying a profile; IFC instances are always implemented against domain-specific schema subsets called Model View Definitions (MVDs). For all three standards, the core component must be implemented by a schema subset. Support for generic object representation is covered by CityGML's Generics extension module and by IFC through the use of IfcClassificationReference, which can be used on all of the IfcProduct specialisations (including IfcProxy); LandInfra currently has no such generics capability. Figure 2 shows the structure, extensibility, and means of sub-setting each schema - the Venn diagram also indicates that the concept of a building lies in the region where the three standards overlap thematically.

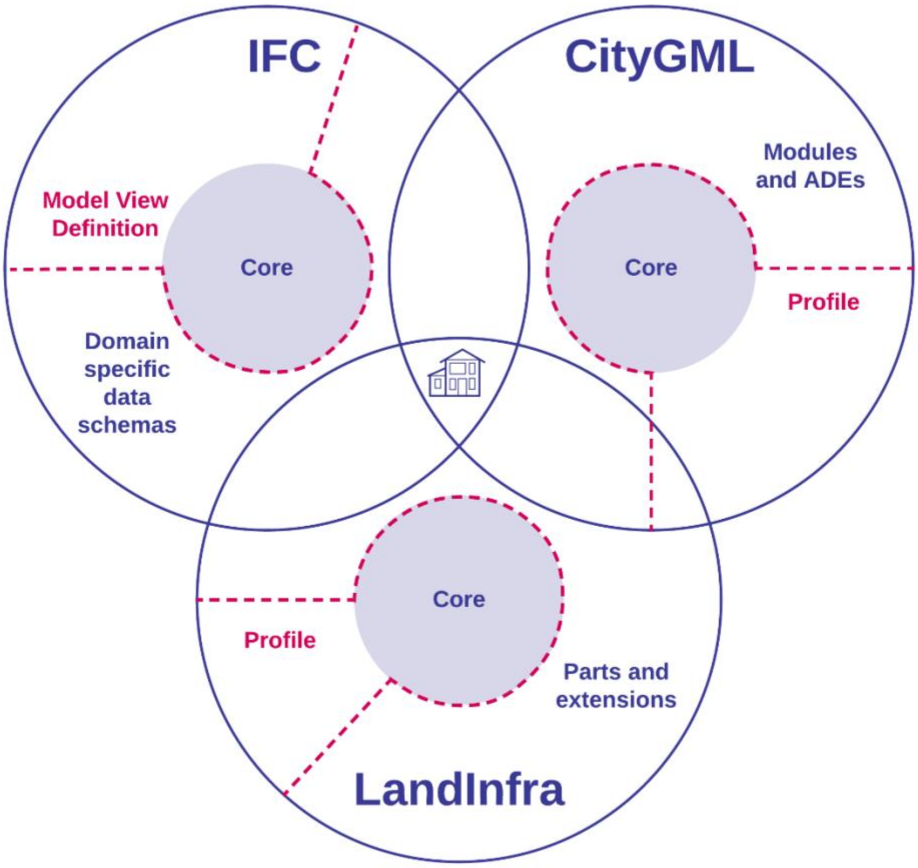

Figure 2 – The three standards differ but overlap in their thematic coverage; as an example, the concept of a building is common to all three. IFC is always subset into Model View Definitions (MVDs) for implementation, whereas the GML-based standards of CityGML and (the InfraGML implementation of) LandInfra can be subset optionally into profiles; for all three, the core of the schema must be implemented. Application Domain Extensions (ADEs) enable anyone to extend the CityGML standard to accommodate more specialist themes;

such extensibility is also possible for the InfraGML implementation of LandInfra. The IFC extensions for domains are part of the buildingSMART governance process and therefore published as formal MVDs.

Conceptualisation and semantics: within their overlapping scopes, the standards differ in how they conceptualise real-world objects. For example, through its spatial structure, IFC allows the representation of a building as an aggregation of parts (such as IfcBuildingStorey elements), which may or may not contain 'spaces' (IfcSpace elements); however, CityGML represents internal building volumes as rooms, and LandInfra represents buildings as features of a facility with only an outer shape and footprint. Semantic inconsistencies add to the disparity, such as how the word ‘feature’ refers to an object in GIS but understood as an embellishment on an object (or element) in BIM; the meaning of data may depend on the context of interpretation. The differences in conceptual modelling of a simple building are depicted in Figure 3. Some conceptual differences can be reconciled; for example, although the meaning of a room in CityGML is different from that of a space in IFC, it might be asserted that particular instances of each refer to the same volume. A mapping from one to the other may then be developed, subject to constraints imposed by differences in geometric representations. The use of a space concept has been proposed for CityGML (alongside the concept of rooms) in order to provide a different means of grouping and querying volumes; a space that is attributed storey could contain members that intersect a horizontal volume, with some members completely enclosed (such as an office space) but others sitting only partially in the storey (such as a lift shaft). Historically, the formalisation of concepts has differed between the standards: from an early stage, concepts in CityGML and LandInfra were modelled in the Universal Modelling Language (UML) while IFC used EXPRESS and EXPRESS-G; however, along with other changes to the standard under version 5, IFC is now also represented in UML. While Figure 2 indicated that the concept of a building is within the thematic remit of all three standards, Figure 3 provides some detail on how object classes in each standard may be used to represent a very simple building instance.

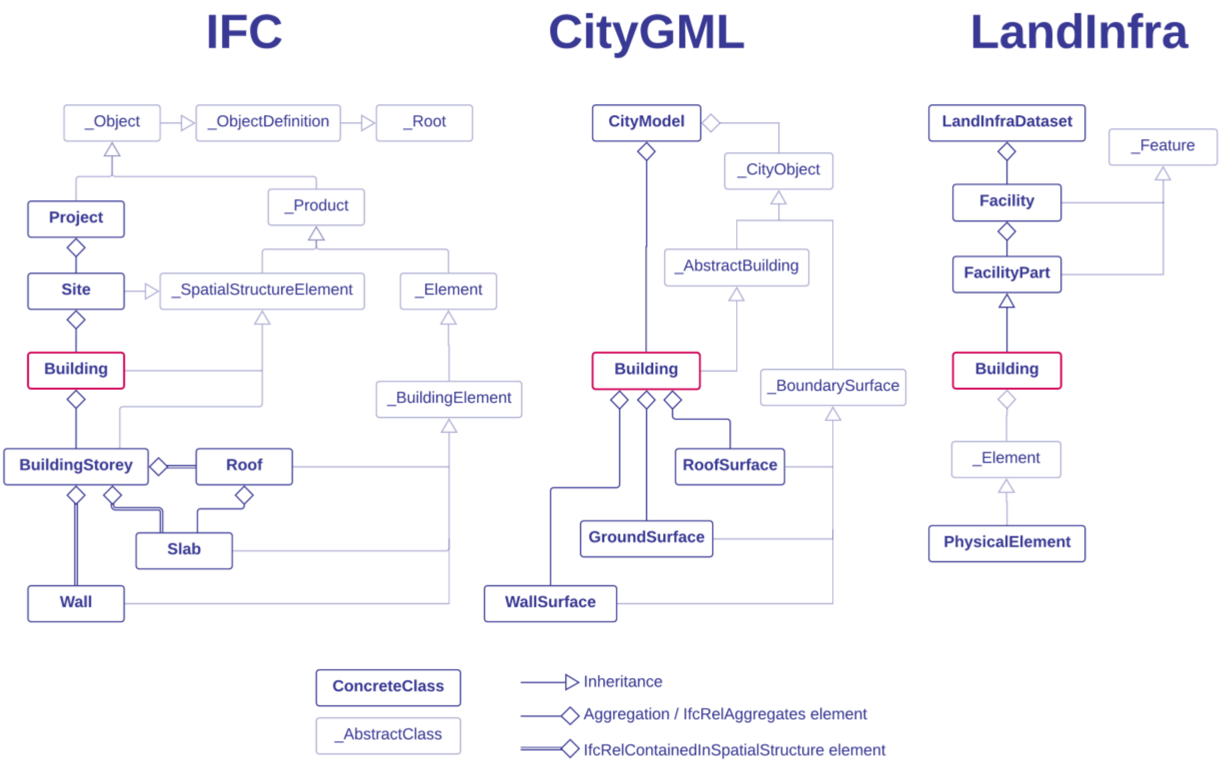

Figure 3 - The concept of a building is represented by all three standards but the detailed structure of the conceptualisation is dissimilar (geometries are not represented in this figure). The UML-like diagrams show possible representations of a very simple building; they are derived from example building instance models that are valid in each of the three standards. The IFC representation is valid for IFC4; in IFC5, a building is a predefined type of facility and a building storey is a specialisation of a facility part.

Coordinate reference systems: the difference in spatial scales and use cases addressed by the standards has led to the employment of two general types of coordinate reference systems (CRS). The IFC schema allows modelling of assets in the high precision required by AEC practices but over areas that are small enough that the subject real-world environment may usually be treated as flat. For this reason, local engineering cartesian CRSs are used for geometry. Although CityGML and LandInfra enable the modelling of finer-scale built assets, both are also concerned with features and collections of features that are large enough to be affected by the Earth's curvature, such that a local engineering CRS alone is insufficient beyond feature geometry; geodetics must be considered and the geographic representations available through GML are used for feature representation or for anchoring features that are represented in a local CRS – in CityGML (and possibly LandInfra), it is more common to use a geodetic CRS for object geometries than anchoring a local cartesian CRS. On the other hand, although it is possible to anchor features geospatially in IFC, this is not yet common practice; when it does take place, reliability in positioning depends on the accuracy of a single coordinate in a geodetic CRS as a reference.

Geometries: the standards are not consistent in their use of geometric representation techniques and this is largely a consequence of the method of data acquisition or generation. The data stored in CityGML are commonly generated from activities such as photogrammetry, laser scanning and transformations from existing 2D landscape models, which are predominantly automated processes. On the other hand, the data represented in IFC (and, to a large extent, LandInfra) are normally generated in manual, interactive design processes. CityGML and LandInfra use only Boundary-Representation (B-Rep) but IFC also supports the parametric modelling techniques of Constructive Solid Geometry (CSG) and Swept Solid, allowing the modeller to choose a representation type that best suits the object and design process. In CSG, objects are constructed using boolean operations on primitive shapes - for example, a tower with a pointed top could be modelled as the union of a cuboid and a pyramid (the height and width of each primitive would be parameters); in B-Rep form, the same tower could be represented by a single volume comprising nine planar surfaces. As a Swept Solid, a pipe could be represented as a circle in the X-Y plane that is extruded (swept) along the Z-axis; the radius of the circle is a parameter that determines the width of the pipe. Such parameterisation can be used for efficient creation and manipulation of more complex (digital) objects whose ultimate form depend on design variables that are not yet finalised; modification requires changes to a few parameters rather than relocating every vertex on a boundary. IFC is increasingly allowing for this parametric modelling paradigm. However, volumes often need to be inferred from observations of real-world surfaces and parametric modelling is not commonly supported by GIS, spatial DBMS and related simulation software. For these reasons, parametric modelling is not supported by CityGML. Although CityGML version 3.0 will support volumetric representations of building elements (in support of transformation from IFC models), these will still need to be represented as B-Reps. Despite LandInfra also not supporting parametrics, given its prescriptive/design role, the inclusion of such a capability has been considered. The absence of a parametric modelling capability is accepted and even preferred by many communities that use CityGML (LandInfra is too young for such an assessment) but this characteristic does make the standard unsuitable for some early-stage urban-scale design tasks for which the standard might otherwise be best suited.

Spatial and network topologies: the standards represent geometric topological relationships differently. In CityGML, topologies such as containment or spatial adjacency are represented implicitly in the elements' boundary representations. Composite objects are constructed using XML's hierarchical graph structure and GML's XLink mechanism. For example, a room could be represented as a composition of several polygon elements that are each defined as a linear ring (a closed loop of straight lines), grouped together as a set of XLinks references. However, IFC would define geometries and the objects that use these geometries, and identify separately any topological relationships explicitly as objectified relationships (such as RelContainedInSpatialStructure). For network topologies, there is more similarity; IFC uses concrete subtypes of abstract IfcRelConnects elements to connect the components of a building's MEP network and, in the CityGML UtilityNetwork ADE, InterFeatureLink elements are used to model connections between features in a utility network.

Progressively detailed geometric-semantic representations: Level of Detail (LOD) is a concept that originates from a need to increase graphics rendering speeds. CityGML uses the LOD concept to formalise and allow multiple and progressively detailed semantic-geometric representations; while roofs and walls may appear in a building at LOD 2, only at LOD 3 and 4 are rooms and furniture modelled. The levels of development (also referred to by the 'LOD' or 'LoD' acronym) in BIM refer to the level of progress through a construction project that is represented by the model and thus do not correspond the geometric LODs in CityGML. The geometric detail stored in an IFC model is generally equivalent to at least the highest CityGML LOD (4) and, if there is a need to convert from IFC to CityGML, a non-trivial mapping process must be developed for the required CityGML LODs. There is no concept of LOD in LandInfra.

Encoding languages: at a more technical level, the predominant encodings that have been used to realise the standards' schemas differ. Although factors such human readability and smaller file sizes can be used to reason a preference for one encoding over another, other factors are likely to have played a larger role in which encoding was selected initially and thus became established. In particular, the availability of technology was a constraint. IFC is the oldest of the three standards; it was conceived as a project in 1994, around the same time that ISO published the initial release of STEP Physical Files (in the EXPRESS modelling language), which had been in development since 1984. The EXPRESS modelling language and instantiating STEP Physical Files proved convenient for modelling the details of buildings, becoming the primary encoding for IFC. The first working draft of XML was released in 1996 and IFC subsequently supported an XML encoding of the schema as ifcXML (an encoding that is now preferred by some users) and a linked data representation in ifcOWL. GML, which is based on XML, arrived in 2000. Importantly, it was GML version 3 (2003) that enabled the representation of 3D objects – development of CityGML began later that same year and its first official release was in 2008. The initial choice of GML for CityGML (thus XSD for schema representation) was influenced by the preference within the geospatial domain for GIS-compatibility or at least the anticipation of geospatial service-based architectures. Having been conceived in 2013 and first released in 2017, LandInfra is much younger than the other two and was designed in the context of their progression and uptake. Although the LandInfra conceptual model is accompanied by the InfraGML encoding (based on GML), it is intended to support alternative encodings, including JSON (which has been implemented as InfraJSON) and EXPRESS. IFC now also supports a wide range of encodings (including OWL and soon JSON), which can be automatically generated from the conceptual UML.

Globally unique identification of real-world objects: format and scope of uniqueness of object identification is also not consistent. IFC enforces globally unique 128-bit number identifiers for all object instances. In contrast, with its dependency on GML and so XML, CityGML requires only that an identifier begins with a letter or underscore and is unique within the scope of the instance document but not necessarily globally. LandInfra allows explicitly for the scope of unique identification of objects to be extended beyond the dataset and made global. However, even in cases where globally unique identifiers (GUIDs) are used throughout all source instance models, individually unique but different identifiers may be used for the same real-world object. This situation is further complicated by capabilities such as multiple representations of a single real-world object in IFC, with each one suitable for a different purpose (architecture or engineering, for example).

3. Challenges to integration

The consequence of the disparities between the standards is that software interoperability and data integration are difficult; they are facilitated by overlapping scope but complications and challenges arise when schemas differ in how they conceptualise real-world objects within these overlapping scopes, how this conceptualisation is formalised and how objects are represented spatially.

Differences in the conceptualisation of identical real-world objects demand subjective intervention to interpret any equivalence or similarity. For some use cases, there is a requirement to map between these concepts and it is not always trivial to find a meaningful correspondence; such an assertion would be necessary for enabling consistent and complete semantic querying of a digital built environment that contains data from such disparate sources. The use of different terms for similar or identical objects – or the same terms for different characteristics – poses an additional barrier. Where concepts are aligned, the lack of convention for unique identification of the same real-world object disallows automatic merging of objects such as infrastructure nodes, hindering the integration of spatial and network topologies.

Inconsistency in the accuracy and method of geolocating objects accounts for some of the integration challenges. Use cases involving spatial querying, three-dimensional visualisation and inference of network topologies across multiple datasets demand a level of coherence in real-world positioning. The use of geodetic CRSs in CityGML allows for easier indexing in GIS software, faster queries in spatial database management systems (DBMS) and easier spatial integration of different datasets (there is the disadvantage that to change the global position of an object requires a change to the position of every geometric feature). The use of a geospatially anchored local cartesian CRSs in IFC means that objects need to be transformed into a geodetic CRS if they are to be integrated spatially with other geolocated objects (for example, from a CityGML model or other geolocated IFC model). Furthermore, it is not common practice for IFC models to make use of an accurate geolocation as an anchor, which can render spatial integration impossible to automate. In particular, the differences in representation of network topologies between the standards – along with inconsistency in the format and scope of real-world object identification – highlights the need to be able to use real-world positions to infer relationships between objects, identify object duplication and resolve conflict or contradiction.

Differences in geometric representations and discrepancies in parametric modelling capabilities also hinder integration. The restriction of CityGML and LandInfra to the Boundary-Representation (B-Rep) method does simplify (to some extent) their implementation – for example, there is no need to deconstruct Constructive Solid Geometry (CSG) trees. However, CSG geometry must then be translated to B-Reps if an IFC model needs to be converted to CityGML or LandInfra. Furthermore, CityGML represents curves as a set of straight lines, which can make conversions from IFC to CityGML in a completely 'lossless' way practically impossible, although often not necessary or even desirable. In general, it is easier and more common to translate from a prescriptive standard to a descriptive one – for example, from IFC to CityGML. Fine detail building features represented in a high-precision local engineering CRS can be mapped geospatially with geometric details abstracted to the coarser resolutions required by city models, and values can be passed for parameters to realise non-parametric states (there is a technological issue with handling and storing extremely high precision coordinates from BIM models in a geospatial CRS). The converse task of enriching construction designs with real-world observation data is frustrated by the dilemma that visible object boundaries are often the only observables, which may be insufficient for the volumetric, parametric representations demanded by architects and construction engineers. Although there is an unambiguous mapping of CSG and Swept Solid geometries to B-Reps, there is potentially an infinite number of options for the converse. Determining the most appropriate means of representing within BIM models the irregular shapes derived from real-world surface observations remains a challenge and topic of discussion.

The standards' broader type of usage and accepted credibility or trustworthiness pose further challenges. The use of IFC as an exchange format alongside working native models (such as Revit files) has been associated with a lack of confidence in the reliability of the format in representing all of the BIM project information necessary for each stage of a construction lifecycle. This might imply a lack of sufficient coverage by the schema or that it is not being used to its full capacity. This issue is not apparent for CityGML, which might be attributed to the common use of CityGML as a native/working modelling schema, lower expectations on data fidelity and that there are fewer implementations (compared with IFC). LandInfra is too young for its uptake to be evaluated. Given the breadth of use cases (including the ability to exchange data, model existing environments and design for the future) there is some concern that user confidence in an integrated digital environment is limited to the level of confidence given to the schema of the least trusted data source.

4. Discussion

It seems reasonable to suggest that there is a consensus for a future of spatial data integration that looks different from how it is today. The situation is complex, and several questions and considerations arise through the discussion of approaches to achieving the desired change.

There is a trend towards the common use of UML across the standards, easing the communication of conceptual designs and reducing the burden involved in developing software for automatic encoding generation. However, there remain semantic disparities between concepts. In order to avoid ambiguous terminology, a shared vocabulary with agreed definitions could be introduced, although this carries the risk of breaking backwards compatibility; alternatively, a shared resource for identifying synonymous terms in different domain vocabularies might be more feasible. There is also a need to discuss whether it is acceptable for the standards to remain predominantly descriptive or prescriptive, or to evolve to accommodate the convergence of the traditionally distinct working domains. Inconsistencies in geolocation continue to present challenges to integration and there remain unsolved problems in reconciling the differences in geometric representations.

In order to understand the merits of proposed solutions for overcoming the challenges faced, it is useful to categorise integration into broad paradigms.

4.1. Integration paradigms

Schema mapping: this is a process that enables an object in one schema to be converted to its equivalent (or nearest match) in another, constituting a complete read-rewrite process. Where there is insufficient overlap, hooks can be engineered using mechanisms such as CityGML ADEs. The GeoBIM ADE from 2009 was one of the earliest examples of this; around ten years later, further efforts were made in the Virtual Singapore project to bring a meaningful subset of IFC concepts into CityGML, and by TU Delft to capture concepts from LandInfra. Any incongruousness between the mapped schemas needs to be reconciled by identifying an acceptable level of equivalence. This single-schema integration paradigm depends on the ability of one schema to represent all of the desired information.

Federation: an alternative paradigm is one in which a software environment is able to simultaneously interpret instances from multiple schemas and offer functionality that operates across the collective dataset; this is particularly relevant given the general trend away from static file-based representation and towards web services that are connected to various static and dynamic data sources. Such multi-schema federation allows for more conflict in conceptualisation. Given the disparity of source schemas, it is worth considering whether the inheritance hierarchy of existing standards are an appropriate way of representing the world. The ‘composition over inheritance’ principle of object-oriented programming could be considered but imposing such a principle is more feasible for the development of a new standard. However, consistent use of geographic CRS and accurate geometric representations are feasibly applicable to existing standards and would enable some integration tasks that make use of spatial data. For example, components of utility networks that are physically connected in the real world but digitally modelled with different schemas – and thus different instance models – could be connected through inferred relationships based on spatial topological relationships such as coincidence or intersection. In other use cases, such as querying a digital model, such precise spatial representation wouldn’t be required: components from IFC, CityGML and LandInfra instances could be pooled in one store and automatically assigned to semantic categories by attributes. The components could then be queried by theme and according to the containment of their approximate centroids within regions of interest. For example, an urban model could be queried for the number of windows within a city precinct, returning components that derive from both IFC and CityGML models. This approach has been demonstrated by TU Munich with the development of the QL4BIM language, by the linked data community using SPARQL on triples of IFC and CityGML, and by TU Delft on a canonical data model. In such an environment, the ability to map across schemas losslessly becomes less important as the objective is to develop integrated digital representations that enable the execution of queries, analyses and visualisations that operate with sufficient fidelity across the digital environment; instead of aiming for losslessness, the objective then is to achieve a level of information recovery that is appropriate for target use cases.

Link referencing: a third paradigm achieves integration through embedding links (such as Uniform Resource Identifiers - URIs) in the primary, working model to components from instances of secondary schemas. The feasibility of such an approach was confirmed in OGC Testbed 4 (2007), in which web services were used for 'merging' CAD, GIS and BIM data. As an example of this paradigm, a building footprint in an urban model might be linked to a BIM model that provides a detailed 3D representation. In CityGML, such a link would be stored as an ExternalReference, which is specified as a URI. Linked data resources allow this capability on a more granular level. Globally unique and common identifiers that persist for such an entity in perpetuity (or at least within its lifecycle) would ensure the dependability of such linking but this demands long-term information management.

4.2. Additional considerations

Effective approaches to integration may involve a combination of paradigms and this choice will likely depend on the domain of application and specific use cases. In many of the fundamental operations that underpin the approaches, a harmonisation of concepts facilitates integration; the ability to return meaningful results from queries across disparate data sources and model interdependencies across federated models depend on the ability to assert some level of commonality in conceptual representations.

The data standards have been developed over many years in response to a range of meaningful, practical use cases and it is neither feasible nor desirable to redesign them from scratch. However, if there is consensus for more harmonisation, the community could look to the joint development of the alignment concept for IFC and LandInfra as a guiding example. The IFC for Infrastructure project used a ‘process analysis’ to identify use cases in support of the design process. The resulting schema extension can be used as a common resource by the linear transportation infrastructures of railways, roads, bridges and tunnels. Traditional 2D horizontal and vertical alignment concepts were extended with a new 3D concept that uses cross-section profiles. The conceptual model was developed in collaboration with OGC and the alignment concept is implemented in LandInfra for both railways and roads. Other concepts could be identified for opportunistic harmonisation, on the understanding that the process is more feasible for new concepts or those that need to be reworked. This would result in an incremental harmonisation of the standards. The potential inclusion of bridge and tunnel concepts in LandInfra could be an opportunity for harmonisation with existing equivalent concepts in IFC and/or CityGML. However, it is pragmatic and beneficial to allow some differences between conceptual representations; any complete, lossless representation of one standard in the schema of another achieves only a different encoding of the same thing and a single, common conceptual schema is unlikely to allow for the different use cases against which the standards were originally developed.

Disparity and conflict between use cases pulls the development of standards in different directions. Irrespective of the approach or approaches that gain consensus, if an unquestionable objective is to minimise the intervention required from a user to achieve a viable and useful level of integration, there is a pressing need to articulate both existing and plausible future use cases against which any approach can be implemented and assessed. It will be important to define the type of conceptual conflicts encountered in a use case and determine the extent to which they are permissible given the required type and depth of integration. The software applications that are used to integrate the data, or within which the integrated data will be used, should also be examined; a closer look at their input requirements and internal digital representations will enable us to better understand and hence address the challenges encountered when bringing together these datasets from disparate data standards. The use of straightforward, non-technical, inclusive language - along with details from specific case studies - will ensure that the communication can be easily digested and consistently interpreted, fostering participation from a wide range of stakeholders and a common understanding of the key problems and objectives.

Broadly speaking, there is a balance to be struck between the freedom afforded by allowing conflicting understandings of identical real-world objects and the ease of integration enabled by harmonisation or abstraction. If harmonisation of concepts by standards organisations is implemented indefinitely, the standards will tend to converge. Equally, if conflicts are left to resolve naturally within the user community, one conceptualisation might eventually out-compete the others as a consensus grows. A consideration with the latter approach is potential bias in user preference towards standards that are supported by software from large, profit-driven organisations; choosing such software is likely to present a lower investment risk but openness of standards is likely to be less interesting to the software provider if cross- or multi-platform interoperability yields more competition. An approach of abstraction of existing standards is likely to be neither feasible nor desirable given the value of (and dependency on) the rich modelling present in each standard. The objective, then, is to devise an approach that allows freedom of real-world conceptualisation, encourages harmonisation where it is beneficial and feasible, and potentially standardises the method of integration itself.

5. Proposed actions

The following are proposed as action points for achieving better interoperability and integration:

- Articulate clearly a set of illustrative use cases in plain, succinct language that ensures they are digestible by a broad audience, basing them on material in existing technical use case documentation. These use cases should include details of the software applications that are commonly used for integration or working with integrated data, including their input requirements and internal representations.

- Derive and make publicly available a shared vocabulary or definition dictionary from terms that are already used in the standards, or a shared resource for identifying synonyms.

- Author a best practice document that recommends the use of three-dimensional georeferencing that is expressed at an appropriate level of precision, defines the level of confidence or accuracy and states the data provenance.

- Devise a system of common unique identifiers for real-world, physical objects that remain constant and exclusive to the objects either in perpetuity or for as long as the objects exist.

- Agree on a collaborative mechanism for opportunistic harmonisation of conceptual representation at thematic overlaps so far as this does not inhibit enrichment and refinement of the schemas as required within their respective domains.

6. References

IFC documentation links:
https://www.buildingsmart.org/standards/bsi-standards/industry-foundation-classes/
https://technical.buildingsmart.org/standards/ifc
https://www.iso.org/standard/70303.html
https://standards.buildingsmart.org/IFC/DEV/IFC4_2/FINAL/HTML/

CityGML documentation links:
https://www.opengeospatial.org/standards/citygml
http://www.citygml.org/

LandInfra documentation links:
https://www.opengeospatial.org/standards/landinfra
https://www.opengeospatial.org/standards/infragml
https://docs.opengeospatial.org/is/15-111r1/15-111r1.html

This document is not intended as an academic publication and a decision was made to omit further referencing.