AUTOMATED GENERATION OF SEMANTIC IFC MODELS FROM LEGACY SEWER NETWORK DATA

Peter Gölzhäuser, Raquel Valles Gomez, Sabine Hartmann, Hendrik Morgenstern, and Katharina Klemt-Albert
Chair and Institute of Construction Management, Digital Engineering and Robotics in Construction, RWTH Aachen University, Aachen, Germany

Abstract

This paper presents an automated workflow for generating georeferenced IFC 4.3 models from legacy data structured according to the DWA-M 150 schema. The DWA-M 150 schema is an XML-based data exchange format for condition monitoring of drainage systems, which is used as a standard by German municipal utilities. The approach consists of three stages: geometry generation from master inventory data, semantic enrichment, and integration of inspection and condition assessment data. The resulting semantically enriched models provide a solid foundation for future digital twin applications in sewer infrastructure management.

Introduction

Germany's drainage and sewerage infrastructure faces the dual challenge of aging assets and necessary expansion (Berger et al., 2020). Furthermore, climate change exacerbates these pressures, with an increased frequency of heavy rain events heightening the risk of flooding (Blöschl et al., 2019). Against this context, the complexity of modern infrastructure assets, coupled with the high cost of labour, makes manual inspection and condition data interpretation increasingly impractical (Tse, 2016). Adopting a more data-driven and automated approach may offer a solution to this problem. However, the data landscape is characterised by fragmentation, often referred to as ''data silos'', which poses a significant obstacle to the adoption of advanced predictive maintenance strategies (Kretschmer et al., 2023). To address this, researchers propose standardized data models that enable both human comprehension and computational reasoning, facilitating seamless integration of heterogeneous data streams into a coherent system view (Du et al., 2023).

This paper presents an automated approach for generating georeferenced IFC 4.3 models from data based on the DWA-M 150 schema, including the visualization of inspection and condition assessment results. By transforming existing municipal sewer data into semantically enriched 3D models, this approach provides the necessary foundation for future digital twin applications.

Background

This section establishes the theoretical framework for the study. It contextualizes the current state of digital sewer data management, contrasting existing practices with the requirements of digital twins and the IFC 4.3 schema. It distinguishes between model generation methods and explains the principles of semantic enrichment to translate structured legacy data into 3D data models.

Digital transformation in sewer infrastructure management

The growing demand for integrated data management drives the construction sector towards the adoption of the *digital twin (DT)*, defined as a dynamic digital representation of a physical asset (buildingSMART International, 2020). This definition can be refined by distinguishing between a *Digital Shadow* (DS)—where data flows automatically from the physical to the digital object—and a full Digital Twin, which includes a bidirectional flow allowing for autonomous control (Ghorbani Bam et al., 2025). To operationalize these concepts, current research proposes a layered reference architecture consisting of Presentation, Platform, Transmission, and Data Acquisition layers (Smarsly et al., 2025). In line with this framework, our study falls within the platform layer, as it focuses on the elements that make up the digital replica, in particular the modeling and data storage elements required for a foundational digital representation.

Currently, sewage networks are typically structured topologically and described using a node-edge model. In this model, nodes represent structural elements such as manholes, inspection chambers and sewage treatment plants, while edges represent connected pipe sections (DWA, 2021). In Germany, the management and standardization of this data are governed by the DWA-M 150 guideline (*Datenaustausch für die Zustandserfassung von Entwässerungssystemen*) (DWA, 2010). In terms of spatial context, Geographic Information Systems (GIS) have become the standard at a macro level. However, they lack the geometric granularity required for component-level maintenance, a task for which Building Information Modeling (BIM) has been proven to be useful (Wang et al., 2019; Valles Gomez et al., 2025).

The role of Building Information Modeling and open standards

The release of IFC 4.3 as an ISO standard in 2024 (ISO 16739-1:2024) marked a milestone by extending the schema to include comprehensive support for infrastructure assets. While core entities like `IfcPipeSegment`, `IfcWasteTerminal` and `IfcDistributionSystem` existed in earlier versions for vertical construction, IFC 4.3 introduces the specialized predefined types and linear referencing capabilities necessary to model civil works such as highway drainage or municipal utilities. This standardisation paves the way for a semantically structured 3D data model. Its open data structures allow for high interoperability, enabling the future integration of sensor data for managing municipal assets throughout their lifecycle (Edmondson et al., 2018).

To bridge the gap between existing infrastructure records and new digital standards, two methodological approaches can be differentiated:

1. **Scan-to-BIM (Reality Capture):** Processes involve capturing spatial data via Terrestrial Laser Scanning (TLS), photogrammetry, or SLAM-based mobile mapping to generate high-density point clouds, followed by a reconstruction process. Recent research has focused on automating the subsequent reconstruction process via Deep Learning architectures for semantic segmentation (Jäkel et al., 2025; Kellner et al., 2025). While effective for visible, aboveground structures such as bridges or facades, the lack of line-of-sight and the logistical constraints of physical access render reality capture and large-scale modeling for extensive underground utility networks highly challenging.

2. **Legacy-to-BIM (Information Reconstruction):** Often characterized as semantic enrichment in recent literature (Belsky et al., 2016; Jiang et al., 2023), reconstructs or augments models from existing datasets. The approach proposed in this paper uses the implicit semantics within legacy records (such as structured XML files like DWA-M 150) to algorithmically generate semantics and geometry. The research challenge lies in translating data attributes into the 3D shapes, properties, and topological relationships required by BIM standards.

While legacy data for sewer systems is often highly available and structured in domain-specific formats (e.g., DWA-M 150), there is currently no direct, standardized mapping to transform this data into the explicit geometry and semantic structure required by the IFC 4.3 standard.

Approaches to geometric and semantic enrichment

The concept of *semantic enrichment (SE)* has emerged as a response to the interoperability gaps inherent in digital construction workflow (Sacks et al., 2017). Addressing this challenge, Belsky et al. introduced ‘‘SeeBIM’’ (Semantic Enrichment Engine for BIM), a prototypical framework designed to interpret the implicit information contained in building models. Their rule-based inference engine utilizes domain expert knowledge—such as the connection constraints between precast concrete elements—encoded into ‘‘IF-THEN’’ logic to deduce new semantic objects and relationships from raw geometry. While highly effective, this approach fundamentally relies on inferring semantics from geometric and topological evidence within standard IFC files (Belsky et al., 2016). The workflow follows a logical path of Geometry → Topology → Semantics, deriving meaning from scanned shapes (Sacks et al., 2017). In the context of this study, however, the challenge is inverted. The source data (DWA-M 150) is semantically rich, but does not include a 3D-model. Consequently, the approach required for legacy sewer data must follow the path of Semantics → Topology → Geometry, utilising the structured attributes in the XML to generate the geometric representation.

This methodological inversion prompts a re-evaluation of the term ‘‘semantic enrichment’’. Bloch refines the prior definition by distinguishing between enrichment and enhancement. In her framework, *semantic enrichment* describes a process using inference engines to derive new information from the model itself (creating new knowledge), whereas *semantic enhancement* refers to the integration of external data sources into the model (linking existing knowledge) (Bloch, 2022). The workflow proposed in this paper occupies a hybrid position between these two concepts: it *enhances* the model by mapping external attributes to IFC property sets, yet it *enriches* the model by algorithmically generating 3D geometry and topological connections that did not previously exist in the 2D source data. Thus, while acknowledging the taxonomic distinction between enrichment and enhancement proposed by Bloch (2022), this paper adopts semantic enrichment as a comprehensive umbrella term to describe the hybrid process of both inferring model geometry and integrating external legacy records.

Research gap

Current research focuses on either digital twin dynamics (e.g., simulation, sensors) or geometric reconstruction (Scan-to-BIM), but lacks standardised methods for the automated semantic translation of legacy data sources into open BIM standards.

Since the DWA-M 150 schema provides rich, structured semantic data, converting these records to IFC 4.3 through a ‘‘Legacy-to-BIM’’ approach offers a pragmatic solution. This methodology is particularly relevant given the proliferation of Machine Learning techniques in both semantic segmentation and enrichment (Schönfelder et al., 2023). However, by their very nature, these approaches entail probabilistic uncertainties (Kellner et al., 2026). In contrast, the use of a well-structured database source enables deterministic, rule-based algorithms to create georeferenced semantic models with high accuracy, thereby circumventing the uncertainty associated with machine learning.

Methodology

This section presents the technical methodology for the automated generation of the IFC model and semantic enrichment from DWA-M 150 data. The proposed approach maintains a separation of geometry and semantics, enabling both processes to run independently of each other. This allows for flexibility in scenarios where geometry originates from alternative sources, such as point clouds resulting from 3D scans, and facilitates regularly updating condition data resulting from new inspections. Figure 1 shows a schematic representation of the conversion pipeline.

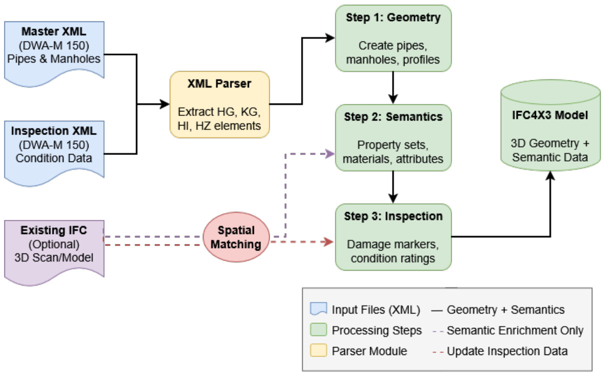

Figure 1: Overview of the conversion pipeline

Data basis and requirements

The process relies on input data structured according to the DWA-M 150 schema, which has been widely adopted by German municipalities and is integrated into most Sewer Information Systems. Although the algorithm presented in this paper was designed specifically for this standard, the underlying approach is transferable and adaptable to other structured data formats.

Within DWA-M 150, each infrastructure element is described by a set of standardized identifiers that are defined as mandatory or optional within the schema. Pipe sections (''Haltungen'' in German) use the identifier prefix `<HG>`, while nodes (''Knoten'' in German) use `<KG>`. The standard defines four element categories:

- **Basic Data** (HG/KG): Designation, material, dimensions, construction year
- **Geometry Data** (GO/GP): Coordinates and spatial information
- **Inspection Data** (HI/KI): Camera survey metadata and observations
- **Condition Data** (HZ/KZ): Damage classifications and assessments

The source file functions as a hierarchical data tree where the root element `<Data>` contains collections of pipes `<HG>` and nodes `<KG>`. Geometry is explicitly defined in `<GO>` (Geometry Objects) subsections, providing Easting, Northing, and Elevation values for each feature point `<GP>`.

Automated geometry generation workflow

The workflow is implemented as an automated converter in Python using IfcOpenShell for IFC 4.3.2.0 generation and NumPy for vector-based geometry calculations. Its geometry module handles coordinate transformations, profile creation, and extrusion operations. The converter also includes a parser module for the DWA-M 150 XML serialisation, which handles all element types (HG, KG, HI, HZ, and HM) and a mappings module for code translations and reference table handling.

Profile definition

The cross-sectional profile of each pipe is determined from the profile type code (`HG305`) and dimensional attributes (`HG306`, `HG307`). The converter supports three profile types:

- **Circular profiles** (code `DN`): Most common, represented as `IfcCircleHollowProfileDef`
- **Egg-shaped profiles** (code `EI`): Typical for combined sewers, created as custom parametric profiles using `IfcArbitraryProfileDefWithVoids` with standard DIN ratios.
- **Rectangular profiles** (code `RE`): Used for larger collectors, represented as `IfcRectangleHollowProfileDef`

Since the wall thickness is not specified in DWA-M 150, a diameter-dependent lookup table based on common precast concrete specifications is used (e.g., 60 mm for DN 150, 75 mm for DN 300–500, up to 220 mm for DN 2200), with linear interpolation for intermediate sizes.

Elevation handling

Missing Z-values are common in legacy datasets. The converter employs linear interpolation when multiple elevation points are known, assumes a flat profile when only one is available, and applies a configurable default value when no elevations exist. To ensure transparency, the converter uses a custom property set, `DWAM150_DataQuality`, to document the assumptions made.

Geometry extrusion

Pipes are modeled as `IfcPipeSegment` entities by extruding the assigned profile along the direction vector between consecutive geometry points. When a pipe section contains more than two geometry points, the converter creates multiple connected linear segments to approximate the polyline path. Manholes and inspection chambers are created as `IfcDistributionChamberElement` elements, with parametric geometry —consisting of a main shaft, an eccentric tapered cone, adjustment rings, and a cover plate— with a forced minimum height of 0.8 meters to prevent geometric degeneration when cover and invert elevations are identical.

Visual quality indicators

To ensure transparency regarding data completeness, color-coding distinguishes elements with complete data from those requiring assumptions:

- **Blue**: Pipes with complete data

- **Red**: Pipes with missing or interpolated values
- **Yellow**: Nodes with complete data
- **Orange**: Nodes with incomplete data

Semantic enrichment framework

The DWA-M 150 schema is optimised for data transport between systems, encoding attributes as 5-character field codes (e.g., HG304 for pipe material) with values referencing internal code tables. For a functional digital twin, this technical encoding must be transformed into human-readable, queryable IFC property sets. The semantics module handles this transformation through code mapping, property set assignment, and spatial matching for data fusion.

Data mapping strategy

The DWA-M 150 standard encodes attributes using a two-level reference system. At the global level, field codes identify attribute types (e.g., KG305 for the manhole material). At the local level, attribute values are numeric codes referencing lookup tables within the XML file. A mapping dictionary translates 5-character field codes into human-readable German property names. For example:

- HG001 → ''Bezeichnung'' (Designation)
- HG304 → ''Material''
- HG305 → ''Profilart'' (Profile Type)
- HZ001 → ''Station'' (Position along pipe)
- HZ014 → ''Schadensklasse'' (Damage Classification)

Semantic values are resolved using a two-stage lookup: the parser first dynamically builds a dictionary from the XML's internal Reference Tables (<RT> elements), then falls back to built-in mappings for codes not defined in the source file. For instance, in the material field HG304, a value of ''10'' is resolved to ''Beton'' (Concrete), ''50'' to ''Steinzeug'' (Vitrified Clay), and ''70'' to ''Kunststoff'' (Plastic).

Spatial matching algorithm

The converter supports two strategies for assigning semantic data to geometry: name-based matching and spatial matching. By default, the converter uses the pipe designation (HG001) to find the corresponding IFC element by name. This works well when the same data source provides both geometry and attribute data. Spatial matching becomes necessary in several scenarios. First, when inspection data comes from a different source than the master data, such as when importing camera survey results from a third-party contractor. Second, when pipe names have changed between data exports or contain inconsistencies. Third, when working with geometry obtained from 3D laser scans, point clouds, or photogrammetry that does not include attribute information. The spatial matching algorithm builds an index tree of all geometric elements in the existing IFC model and calculates the geometric centroid, typically the midpoint between start and end points for pipe sections. First, the algorithm queries the spatial tree at the exact centroid location. If exactly one element intersects that point, a match is found. If no elements are found, the search radius expands exponentially—doubling from an initial radius of 10 cm (0.1, 0.2, 0.4 m, etc.) up to a configurable maximum (default: 5 m) to ensure that coordinate differences due to rounding or survey accuracy do not prevent valid matches. When multiple elements fall within the search radius, the match is considered ambiguous, and no automatic association is made. The converter reports these ambiguous cases so that operators can resolve them manually. The centroid-based approach was chosen for its simplicity. Alternative strategies such as bounding-box overlap, direction-vector similarity, or topological connectivity matching were not implemented in the current version, as the centroid approach proved sufficient for straight pipe segments in the test dataset. The limitations of this algorithm for curved segments and dense junction areas are discussed further in Section 4.5.

Property set assignment

Once a match is found, the attributes are assigned to the IFC object using custom property sets prefixed with DWA_. Custom property sets are needed because the IFC 4.3 standard does not define property sets that fully accommodate the DWA-M 150 domain model—for example, Pset_Condition provides only a single overall assessment field, whereas DWA requires a triple classification (tightness, structural stability, and operational reliability) with per-defect station coding. To ensure both usability and data lineage, two distinct property sets are created for each data category: one includes decoded values for end users, while the *_raw version preserves the original DWA codes for traceability and automated validation. For the static master data (''Stammdaten'' in German), which includes basic and geometry data elements, the following property sets are created:

- DWA_Stammdaten: Contains the decoded attribute values (e.g., Material: ''Concrete''). This is intended for end-users in BIM viewers.

- DWA_Stammdaten_raw: Preserves the original DWA code values (e.g., Material: ''10'').

For inspection and condition data, additional property sets are created following the same approach:

- DWA_Inspektionsdaten / ..._raw: Inspection metadata.
- DWA_Zustandsdaten / ..._raw: Condition observation data.
- DWA_DamageReference: Links damage markers to parent pipes with station, classification, and a human-readable damage description.

Inspection data integration

Beyond static master data, the converter supports inspection records <HI> and condition observations <HZ>. In

DWA-M 150, damage observations are recorded as linear stations — distances in meters from the inspection start point (typically the upstream manhole). The HZ001 field stores this 1D station value, which is converted to a 3D world coordinate by interpolating along the pipe's direction vector. Individual damage observations are visualized as 3D ring-shaped color-coded markers, as shown in Figure 2, according to the German condition classification:

- **Class 0 (Green):** No damage / informational
- **Class 1 (Red):** Immediate action required
- **Class 2 (Orange):** Short-term action needed
- **Class 3 (Yellow):** Medium-term action
- **Class 4 (Grey):** Long-term monitoring

This enables operators to immediately identify critical sections requiring maintenance intervention.

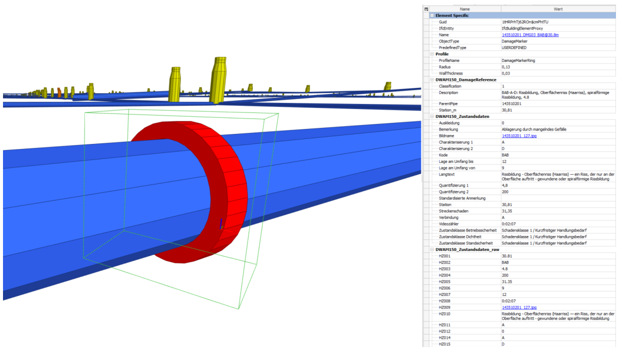

Figure 2: Damage Marker shown in BIMvision

Results and discussion

The workflow was validated using a combination of real-world and synthetic datasets. Real municipal master data from the German municipality of Alsdorf was used to test geometry generation and semantic enrichment with static master data. Although name-based matching is the primary method for data integration from the same source, the spatial matching algorithm was also tested using this dataset. Synthetic inspection data was generated to test the inspection data integration, covering the full range of damage classifications and several station positions across the pipes. The following sections evaluate geometric accuracy, semantic data verification, inspection data integration, and system performance, followed by a discussion of current limitations.

Geometric model accuracy

Georeferencing accuracy was validated by overlaying the generated IFC model on a Digital Terrain Model (DTM) created in Autodesk Civil 3D. The DTM was constructed using GeoTIFF elevation data obtained from the state geodata portal of North Rhine-Westphalia (Geoportal NRW), and orthophotos as visual overlay, referenced to the ETRS89 / UTM zone 32N coordinate system (EPSG:25832). The unmodified IFC file was then superimposed on the DTM using both Desite and Autodesk Infraworks, as illustrated in Figure 3. Coordinates from several pipe elements along the outer boundary of the network were extracted from the IFC file and from the combined model and verified against the original DWA-M 150 source data. This comparison confirmed that UTM coordinates were correctly preserved in the generated model.

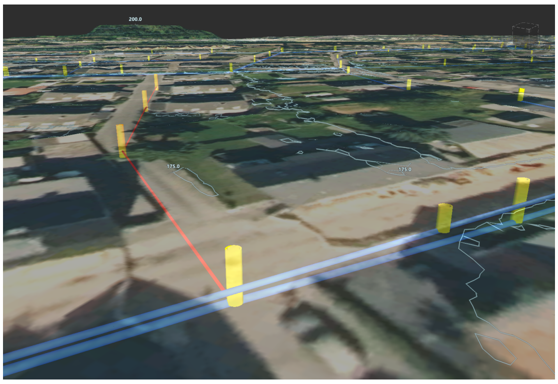

Figure 3: Coordination Model in Autodesk InfraWorks

Semantic data verification

The semantic enrichment was validated using real-world municipal datasets from the German municipality of Alsdorf, containing 701 pipe sections and 702 nodes. Both matching strategies were tested independently. Since both geometry and master data originated from the same DWA-M 150 source and were consistent, name-based matching achieved a 100% success rate, linking all 701 pipes and 702 nodes to their corresponding semantic attributes. Spatial matching achieved a success rate of 97.6% for pipes (684 of 701 matched), with 17 pipes (2.4%) remaining unmatched. 14 of these 17 pipes were curved segments with more than two geometry points, where the centroid-based matching fails due to geometric segmentation. The remaining three pipes were unmatched due to overlapping geometries at complex junction points, where multiple pipes shared near-identical centroids. For nodes, spatial matching achieved 99.6% accuracy (699 of 702). The resulting IFC models were inspected in BIMvision, where the dual property set structure enabled visual verification of key properties (such as material types), while retaining the raw data for quality checks.

Inspection data integration

To validate the inspection data integration, synthetic inspection data was generated following the official DWA-M 150 example file structure. The synthetic dataset was created by selecting 50 pipes from the Alsdorf master data and generating inspection records with DWA-M 150 element hierarchy: pipe sections (<HG>) with geometry objects (<GO>/<GP>), inspection headers (<HI>), and condition observations (<HZ>). The dataset contains 251 condition observations distributed across all five damage classifications. The converter successfully parsed all inspection records and generated correctly positioned damage markers with appropriate severity color-coding. Linear station values across the pipes (HZ001) were correctly transformed to world coordinates via linear interpolation. The precision is limited only by the input coordinate resolution (typically millimetres in UTM) and the station value precision from the source data. However, the matching algorithms could not be tested using the synthetic data, as it was derived directly from the master data with identical pipe names and coordinates, resulting in a 100% matching rate.

An alternative validation method could include introducing controlled perturbations, like coordinate offsets (e.g., 0.5–2.0 m), renaming pipe identifiers to simulate naming inconsistencies between data sources, or merging inspection records from a genuinely independent data export. As no more heterogeneous datasets were available at the time of testing, this remains a limitation, and future work includes testing with real inspection data from third-party contractors, or alternatively, introducing synthetic perturbations to simulate real-world data inconsistencies.

System performance and quality assessment

The developed workflow successfully achieved its primary objective of automated generation and semantic enrichment of 3D sewer models from DWA-M 150 master data. The resulting model features a complete geometric representation of pipes and nodes, correct georeferencing, and comprehensive semantic information based on the DWA-M 150 standard embedded in property sets. This demonstrates the feasibility of automated, standard-based data transfer from DWA-M 150 to IFC using open-source tools such as `IfcOpenShell`. In terms of quality, the converter captured all objects from the source data, provided that the input data adhered to the standard. Semantic consistency was verified by checking that the attributes matched the DWA identifiers, with the raw values preserved in a separate property set to mitigate data loss from missing reference table entries in the source. The output adheres to IFC 4.3.2.0 and was validated using the buildingSMART Validation Service. The models were additionally loaded in BIMvision and Blender (Bonsai) to confirm practical interoperability. Performance was satisfactory, with geometry generation completing in approximately 15 seconds and semantic enrichment in approximately 18 seconds (name-based) or 14 seconds (spatial) on a workstation equipped with an AMD Ryzen 9 8945HS processor and 32 GB RAM.

Limitations and challenges

Regarding the geometric model accuracy, the system faces challenges primarily related to source data quality, such as missing height values, which necessitated defaults. The manholes are represented as simplified parametric geometries, consisting of cylindrical elements, a tapered cone, and a cover plate. This representation does not account for asymmetric bases, varying wall thicknesses, or non-circular shaft cross-sections. A realistic representation of the nodes would be more demanding. In DWA-M 150, curved pipes are represented by three or more geometry points. During IFC generation, these are approximated as multiple straight segments. While this linear representation facilitates the centroid-based matching algorithm, arc-based swept solids could represent curved pipe alignments more smoothly in future iterations. For the semantic enrichment, the centroid-based spatial matching algorithm shows limitations when handling curved pipe sections. The spatial matching algorithm calculates the centroid from all original geometry points, but this centroid may not fall within any single IFC segment's bounding box. It should be noted that the DWA-M 150 schema only provides discrete coordinate points, making linear approximation the only deterministic geometric representation; fitting arc or spline curves would require additional assumptions about curvature that are not supported by the source data. Matching precision could be improved by calculating separate centroids for each segment of curved pipes or by adopting alternative geometric matching strategies. As more diverse datasets become available in the future—including geometry from laser scans, where curved representations may exist—alternative matching strategies for this specific case could also be evaluated.

Regarding validation, the reliance on synthetic data means that the matching algorithms remain untested for inspection and condition data. The current implementation can therefore only be interpreted as a proof-of-concept for the pipeline architecture, with the matching algorithm requiring further validation as real-world datasets become available. Additionally, specific sewer properties like ''Flood Area'' lack native IFC entities, requiring mapping to IFC spatial entities (eg. `IfcSpatialZone` or `IfcGeographicElement`), potentially reducing interoperability with downstream analysis tools.

Comparison with existing approaches

Compared to proprietary Sewer Information Systems and transformation workflows, this open-source approach offers significant advantages in transparency and flexibility. The separation of geometry generation from semantic enrichment allows for the enrichment of any valid IFC geometry, such as that derived from 3D scans, provided the coordinates match. This makes the workflow particularly suitable for research and development contexts, where adaptability is required. Regarding transferability to other infrastructure domains, the workflow components differ in their degree of generalizability. The geometry generation engine (profile extrusion, coordinate transformation, elevation handling) and the spatial matching algorithm can be reused directly for any node-edge network topology. In contrast, the DWA-M 150 parser, the field code mapping dictionaries, and the reference table resolution logic are domain-specific and would require replacement when adapting the workflow to other standards (such as DVGW for gas and water networks). The IFC entity mapping (e.g., which DWA element types correspond to which IFC classes) is also domain-specific but follows a pattern that can be systematically replicated for other schemas.

Conclusion and outlook

This section consolidates the workflow's main contributions and outlines future development directions.

Key achievements

The study successfully developed a two-stage automated workflow that translates DWA-M 150 sewer infrastructure data into a fully structured, georeferenced, and semantically enriched IFC 4.3 model. This work validates the feasibility of bridging traditional Sewer Information Systems (KIS) and Building Information Modeling (BIM) using open standards (IFC, DWA-M 150).

Key technical achievements include decoupling geometric extrusion from semantic enrichment to enable data integration from heterogeneous sources. To facilitate this process, the study developed and evaluated name- and centroid-based spatial matching algorithms, achieving integration accuracies of over 97% for the spatial matching of pipe segments on real-world municipal datasets. Furthermore, the workflow effectively dismantles existing data silos by translating specific DWA-M 150 attributes into standardised, human-readable IFC property sets, while retaining the original codes for algorithmic validation. Finally, this approach has been extended to operational condition monitoring by translating 1D linear inspection stations into visual, colour-coded 3D damage markers. Having been validated against official buildingSMART configuration rules and coordinated with municipal digital terrain models, this Legacy-to-BIM methodology establishes a tangible, standard-compliant foundation for future predictive digital twin applications in underground infrastructure.

Future improvements

Technical improvements could address geometric and topological limitations. Currently, pipes are linearly simplified; integrating OpenCASCADE/STEP via `IfcOpenShell` would allow for complex profiles and curved segments. Connection geometry between pipes and nodes could also be automated for cleaner transitions. The spatial matching algorithm could be extended with directional similarity scoring and topological connectivity analysis to resolve ambiguous cases. To support hydraulic simulations, semantic flow definitions could be implemented using the IFC concept of ''Port Nesting''. Furthermore, hardcoded defaults for missing data could be replaced with dynamic calculations based on slope or rule-based plausibility checks. For inspection data, integration with the BIM Collaboration Format (BCF) would enable linking damage observations to external documentation such as inspection photographs and video frames. However, BCF currently lacks native fields for precise damage coordinates, requiring either format extensions or hybrid approaches.

Standards and digital twin potential

While the current approach aligns with DWA-M 150, the upcoming DWA-M 145-3 standard (incorporating GML for coordinates) will require adaptations in XML parsing logic, though the general methodology remains valid. The enriched IFC model serves as a foundation for a digital twin, enabling flood prevention simulations and integration with sensors for predictive maintenance. While developed for wastewater infrastructure, the methodology is transferable to other utility networks characterized by similar data structures: water distribution, gas networks, and district heating systems all employ node-edge topologies with standardized attribute schemas.

Acknowledgments

This research is part of the *KaSyTwin* project funded by the Federal Ministry of Transport (BMV) within the mFUND program under grant 01FV2059A.
The authors would like to thank *Regionetz GmbH* in Aachen, the *City of Alsdorf*, and *Stadtentwässerungsbetriebe Köln* for providing data and supporting the research efforts.

References

Belsky, M., Sacks, R., and Brilakis, I. (2016). Semantic enrichment for building information modeling. Computer-Aided Civil and Infrastructure Engineering, 31(4):261–274.

Berger, C., Falk, C., Hetzel, F., Pinnekamp, J., Ruppelt, J., Schleiffer, P., and Schmitt, J. (2020). Zustand der kanalisation in deutschland: Ergebnisse der dwa-umfrage 2020. KA Korrespondenz Abwasser, Abfall, 67(12/2020).

Bloch, T. (2022). Connecting research on semantic enrichment of bim - review of approaches, methods and possible applications. Journal of Information Technology in Construction, 27:416–440.

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K., and Živković, N. (2019). Changing climate both increases and decreases european river floods. Nature, 573(7772):108–111.

buildingSMART International (2020). Enabling an ecosystem of digital twins.

Du, H., Wei, L., Dimitrova, V., Magee, D., Clarke, B., Collins, R., Entwisle, D., Eskandari Torbaghan, M., Curioni, G., Stirling, R., Reeves, H., and Cohn, A. G. (2023). City infrastructure ontologies. Computers, Environment and Urban Systems, 104:101991.

DWA, editor (2010). Datenaustauschformat für die Zustandserfassung von Entwässerungssystemen. Number 150 in DWA-Regelwerk Merkblatt DWA-M. DWA, Hennef, april 2010 edition.

DWA, editor (2021). Anforderungen an ein Datenmodell und Schnittstelle (Entwurf). Number DWA-M 145-3 in Kanalinformationssysteme. Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall, Hennef (Sieg), entwurf, september 2021 edition.

Edmondson, V., Cerny, M., Lim, M., Gledson, B., Lockley, S., and Woodward, J. (2018). A smart sewer asset information model to enable an 'internet of things' for operational wastewater management. Automation in Construction, 91:193–205.

Ghorbani Bam, P., Rezaei, N., Roubanis, A., Austin, D., Austin, E., Tarroja, B., Takacs, I., Villez, K., and Rosso, D. (2025). Digital twin applications in the water sector: A review. Water, 17(20):2957.

Jäkel, J.-I., Heinlein, E., Gölzhäuser, P., Kellner, M., Klemt-Albert, K., and Reiterer, A. (2025). How to digitise bridge structures—a systematic review of the status quo for creating digital bim models of existing bridge structures in the operational phase. Infrastructures, 10(3):47.

Jiang, S., Feng, X., Zhang, B., and Shi, J. (2023). Semantic enrichment for bim: Enabling technologies and applications. Advanced Engineering Informatics, 56:101961.

Kellner, M., König, T., Jäkel, J.-I., Klemt-Albert, K., and Reiterer, A. (2025). 3d bridge segmentation using semi-supervised domain adaptation. Automation in Construction, 172:106021.

Kellner, M., Schmitt, A., and Reiterer, A. (2026). Automatic generation of 3D bridge models from 3D point clouds. Results in Engineering, 29:109532.

Kretschmer, F., Franziskowski, S., Huber, F., and Ertl, T. (2023). Chances and barriers of building information modelling in wastewater management. Water science and technology : a journal of the International Association on Water Pollution Research, 87(7):1630–1642.

Sacks, R., Ma, L., Yosef, R., Borrmann, A., Daum, S., and Kattel, U. (2017). Semantic enrichment for building information modeling: Procedure for compiling inference rules and operators for complex geometry. Journal of Computing in Civil Engineering, 31(6).

Schönfelder, P., Aziz, A., Faltin, B., and König, M. (2023). Automating the retrospective generation of As-is BIM models using machine learning. Automation in Construction, 152:104937.

Smarsly, K., Ahmad, M. E., Peralta, P., Al-Zuriqat, T., Al-Nasser, H., Dragos, K., and Geck, C. C. (2025). Digital Twins, Architectures, and Elements in Civil Engineering – A Multivocal Literature Review. In Francis, A., Miresco, E., and Melhado, S., editors, Advances in Information Technology in Civil and Building Engineering, volume 628, pages 507–518. Springer Nature Switzerland.

Tse, J. H. (2016). Challenges of Adopting Knowledge-based Building Information Modeling for E&M Asset Management Supplemented with Mobile Solutions - A Case Study in Public Sewage Pumping Facilities. American Journal of Operations Management and Information Systems, 1:17–33.

Valles Gomez, R., Jamora, J. L., Gölzhäuser, P. J., Hartmann, S. V., Ravandi, A. B., Morgenstern, H. W. L., and Klemt-Albert, K. (2025). An ontology-based approach for sewer infrastructure management. In Tagungsband des 36. Forum Bauinformatik : 24.-26. September 2025. RWTH Aachen University.

Wang, H., Pan, Y., and Luo, X. (2019). Integrating BIM and GIS in green building certification: A building lifecycle perspective. Automation in Construction, 103:41–51.