

**Digital, autonomous, Intelligent and Synchronous system for
Continuous identification, Optimization and Value Extraction of
Resources from the end-of-use built environment**



DISCOVER

D3.1 Annotated Datasets

**WP 3. Rapid prediction of materials and components
(M7 – M31)**

WP Leader: Tecnalia

Submission date: 31 August, 2025



**Funded by
the European Union**

The DISCOVER (GA 101129909) project is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Project name	Digital, autonomous, Intelligent and Synchronous system for Continuous identification, Optimization and Value Extraction of Resources from the end-of-use built environment
Grant Agreement number	101129909
Funding scheme	Horizon Europe
Project Acronym	DISCOVER
Project starting date	01/06/2024
Project duration	48 months
Deliverable number	3.1
Deliverable title	Annotated datasets
Deliverable version	V1
Work Package number	3
Work Package title	Rapid prediction of materials and components
Due date of delivery	31/08/2025
Actual date of delivery	31/08/2025
Dissemination level	PU - Public
Type	R - Document, report ®
Editor(s)	David Garcia Estevez (TECNALIA), María Jose Lopez Osa (TECNALIA), David Ciro Sierra Garcia (TECNALIA), Ana Isabel Torre Bastida (TECNALIA), Jon Aguirre Usandizaga (TECNALIA), Lander Bonilla Viana (TECNALIA).
Contributor(s)	Alba Perez Gracia (UPC), Muhammad Zain Bashir (UPC), Yeray Navarri Soler (UPC), Lluís Bonet Ortuño (UPC), Francisco Di Maio (TU Delft), Yongli Wu (TU Delft), Mohanad Abukmeil (TU Delft), Rahul Tomar (DTT), Manuel Jungmann (DTT).
Reviewer(s)	Alba Perez Gracia (UPC), Gvantsa Jichoshvili (UPC).
Rights	DISCOVER consortium

Confidentiality	
Does this report contain confidential information?	Yes <input type="checkbox"/> No <input checked="" type="checkbox"/>
Is the report restricted to a specific group?	Yes <input type="checkbox"/> No <input checked="" type="checkbox"/>

Document history

Version	Date	Beneficiary	Description
0.1	15.05.2025	TECNALIA	Initial ToC
0.2	11.07.2025	TECNALIA	Preliminary version of the sections
0.3	24.07.2025	TECNALIA, DTT	New sections. First complete version before internal review
1.0	31.08.2025	TECNALIA, UPC	After final review

List of abbreviations

Abbreviation	Meaning
API	Application Programming Interface
BIM	Building Information Modeling
CAD	Computer-Aided Design
CCPA	California Consumer Privacy Act
DMP	Data Management Plan
ETL	Extract, Transform, Load
FAIR	Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
SQL	Structured Query Language

List of tables

Table 1. BIM maturity levels.....	8
Table 2. Data Governance key elements and summary description of related activities.....	11
Table 3. Summary of ETL phases in DISCOVER.....	17
Table 4. Comparison of orchestrator software options.	27
Table 5. Comparison of open source relational databases.	27
Table 6. Comparison of FTP client and server software.....	28

List of figures

Figure 1. The DAMA DMBOK wheel. DAMA International.....	10
Figure 2. Key elements of Data Governance. Databricks.....	11
Figure 3. DISCOVER Data Governance main steps.....	13
Figure 4. Governance Roles and Responsibilities. Robert S. Seiner, KIK Consulting.....	14
Figure 5. ETL process. Informatica.com.	15
Figure 6. ETL and Data Governance processes.....	16
Figure 7. Main phases of DISCOVER BIM 4.0 generation approach.....	17
Figure 8. Example of Image Metadata JSON.....	21
Figure 9. ETL General architecture.....	26
Figure 10. Cloud components of the ETL system.	28
Figure 11. Conceptual Design of the Use Cases infrastructure.	29
Figure 12. Instance-level configuration of the system.....	30

Table of contents

About the DISCOVER project	4
Abstract	5
Introduction.....	6
Premises and State of The Art.....	6
1.1. Data Governance: Concept of Data Pipelines and Data Products	7
1.2. BIM model – Annotations.....	8
1.2.1. BIM Maturity Levels: From CAD to BIM 4.0.....	8
1.2.2. The Need for Data Governance and Enrichment in BIM 4.0.....	9
1.3. Data Governance methodology.....	9
1.3.1. Preprocessing, Quality Control, and Metadata Annotation of Images:.....	13
1.3.2. Generation of the Base IFC Model from Point Clouds:.....	13
1.3.3. Enrichment of the IFC Model with Metadata and Contextual Image Information:.....	14
1.4. Phases, roles and responsibilities.....	14
1.5. Phase 1: Metadata generation ETL.....	18
1.5.1. Main artifacts (Information to be gathered from partners in an Excel sheet)	18
1.5.2. Data sources – origin of data	19
1.5.3. Transformations.....	20
1.5.4. Final Data Product - Minimum set of metadata.....	20
1.6. Phase 2: IFC Generation.....	21
1.6.1. Point Cloud Data Preparation Protocol for BIM Development	22
1.7. Phase 3: IFC Enrichment.....	23
1.8. Reference Architecture: Implementation design	24
1.9. System Requirements.....	24
1.10. Functionalities and components	25
1.11. Technology stack.....	26
1.12. Infrastructure.....	28
1.12.1. ETL testing environment description	29
1.12.2. Instance Descriptions	30
1.12.3. Advantages of Using OpenStack for Deployment.....	31
Conclusions	32
References.....	33

About the DISCOVER project

DISCOVER intends to develop an autonomous, synchronous, continuous and intelligent identification and data analysis system for materials and products in existing end-of-life built works. The proposed approach will provide key stakeholders, including academia research performers, along with construction industry representatives, with data-driven insights to make deconstruction more efficient, optimise the use of resources, improve the environmental footprints and enhance the circularity of construction and demolition, unlocking the potential of end-of-life built works, which will become material banks. The expected outcomes include an autonomous robotic platform coupled with continuous identification tools to scan built works and provide synchronous quantitative and qualitative data from different materials, including complex and concealed elements. Artificial intelligence algorithms will allow a rapid analysis of the properties and characteristics of components, and feed the automated scan-to-BIM model creation. The multi-dimensional BIM, including selective demolition processes, labour productivity, and technical planning, will become a Digital Twin of the demolition site optimised by social, economic, and environmental multi-criteria assessments. This approach will highly contribute to increase significantly the supply of traceable and sustainable construction materials and products to enhance their wider market acceptance, following the waste hierarchy. The social impacts of digital transformation in the construction sector will be considered, and new professional development tools for the relevant stakeholders will be proposed. The system will be tested in four different real demolition sites (Spain, Portugal, Poland and Belgium), offering a complete range of built work typologies and wide geographical coverage to demonstrate the replicability potential of DISCOVER, increasing the project dissemination capacity and awareness among the construction sector.

Abstract

This deliverable defines the data governance methodology and technical infrastructure underpinning the DISCOVER project's approach to generating and managing annotated datasets for BIM 4.0. The central goal is to support the creation of semantically enriched, interoperable data products derived from multimodal sources (images, point clouds, sensors) to enable advanced applications such as predictive maintenance and asset lifecycle optimization.

The governance strategy is built around ETL (Extract, Transform, Load) pipelines organized in three phases:

- **Metadata Generation & Image Annotation:** Led by Tecnia, this phase extracts structured metadata from visual datasets (e.g., RGB, XRF) provided by partners (TUD, UPC) in WP1 and WP2, preparing them for downstream integration. Key outputs include JSON-formatted image metadata with location, time, and device data.
- **IFC Generation from Point Clouds:** Led by DTT, raw point clouds from Lidar and photogrammetry are transformed into geometric models in the IFC format. A robust protocol ensures data quality, spatial referencing, and adequate density for Level of Development (LOD) 200 BIM models.
- **IFC Enrichment:** Metadata and contextual image information are spatially and temporally matched with IFC elements to produce a fully enriched BIM 4.0 model. This integration supports semantic querying and interoperability across use cases.

A modular reference architecture supports deployment flexibility. It includes a cloud-based stack using FastAPI, Apache Airflow, MongoDB, PostgreSQL, and FileZilla, and is virtualized via OpenStack. Each use case operates a tailored pipeline instance with shared governance principles and templates. The infrastructure ensures scalability, traceability, and compliance with FAIR and GDPR requirements.

While real datasets are still in development, this deliverable provides a detailed methodological and technical blueprint to ensure data consistency, quality, and reuse across the project lifecycle.

Introduction

The primary purpose of this document is to define the data governance methodology and technological foundations that will support the generation, annotation, and management of datasets within the DISCOVER project. It outlines the conceptual and technical framework necessary to ensure that data collected from various sources—such as images, point clouds, and sensor readings—can be transformed into high-quality, interoperable data products aligned with the principles of BIM 4.0.

The document is structured into several key sections. It begins with a review of the state of the art in data governance and BIM methodologies, followed by a detailed description of the proposed governance strategy, including roles, responsibilities, and the definition of data products. Subsequent sections describe the ETL (Extract, Transform, Load) pipelines, the types of data sources involved, the transformations to be applied, and the final data products expected. The document also presents the reference technological architecture, system requirements, and infrastructure components necessary to implement the proposed methodology.

The main objective of this deliverable is to provide a shared foundation for the implementation of data governance practices across all use cases in the project. It describes the activities required to ensure data quality, traceability, and semantic enrichment, while also offering a flexible architecture that can be adapted to the specific needs of each use case. Although this document does not include real datasets—since data acquisition is still ongoing—it establishes the methodological and technical groundwork for their future integration.

Finally, this deliverable aligns with the project's Data Management Plan (DMP) and adheres to the FAIR data principles (Findable, Accessible, Interoperable, Reusable). Special attention is given to security and privacy aspects, ensuring that all data handling processes comply with GDPR and other relevant regulations. The methodology described herein has been developed under the leadership of Tecnalia, in collaboration with project partners, and will serve as a reference for future implementation and refinement.

Premises and State of The Art

This section outlines the foundational concepts and current landscape that inform the data governance strategy proposed in this deliverable. It provides a review of relevant methodologies, technologies, and standards—particularly in the context of BIM and multimodal data integration—that serve as the basis for the DISCOVER project's approach.

1.1. Data Governance: Concept of Data Pipelines and Data Products

In BIM ecosystems that integrate multimodal data—such as images, point clouds, sensor data, and semantic models—data governance is not just a best practice, but a necessity. It ensures that data are accurate, consistent, and interoperable across disciplines and systems. In the DISCOVER project, where diverse data types are collected and processed by multiple stakeholders, governance provides the foundation for trust, traceability, and collaboration. Without a clear governance framework, the integration of spatial, temporal, and semantic data becomes error-prone and fragmented, undermining the reliability of the digital twin or BIM 4.0 model.

One of the main technological challenges in this context is the harmonization of heterogeneous data formats and standards. Aligning metadata schemas, coordinating systems, and data quality thresholds across different sources requires robust validation mechanisms and shared conceptualizations. Additionally, ensuring data lineage—tracking the origin and transformation history of each dataset—is essential for auditability and long-term maintenance of the digital asset.

A central concept in modern data governance is the data product—a curated, reusable dataset designed to deliver value across multiple use cases. In DISCOVER, examples of data products include:

- Annotated image datasets enriched with spatial-temporal metadata;
- IFC models generated from point clouds;
- Semantically enriched BIM models combining geometry and contextual data.

These products are not merely outputs of data processing; they are strategic assets that enable applications such as automated inspections, progress monitoring, energy analysis, and predictive maintenance. By treating data as a product, the project ensures that each dataset is designed with usability, quality, and lifecycle in mind, making it a reliable and reusable component of the digital representation of buildings and infrastructure.

However, building high-quality data products presents several challenges. One of the most critical is ensuring semantic consistency across datasets—especially when combining data from different domains (e.g., visual, geometric, temporal). Another is the versioning and lifecycle management of data products, which must evolve alongside the physical asset while maintaining backward compatibility and traceability.

To implement data governance at scale and ensure the consistent delivery of data products, the DISCOVER relies on ETL (Extract, Transform, Load) pipelines. These pipelines automate the flow of data from raw acquisition (e.g., image capture, point cloud scanning) through transformation (e.g., annotation, format conversion, semantic mapping) to structured outputs stored in repositories. For example, transformation raw

images into annotated metadata-rich JSON files; or conversion point clouds into IFC models; and merging these into enriched BIM 4.0 representations.

The automation provided by ETL pipelines reduces manual effort, enforces governance rules, and ensures repeatability and scalability. Yet, designing these pipelines is not trivial. Key technical challenges include handling data volume and velocity, especially when dealing with high-resolution imagery or dense point clouds; ensuring fault tolerance and recovery in distributed processing environments; and maintaining synchronization between pipelines that operate on interdependent datasets. Addressing these challenges requires a combination of robust architecture, modular design, and continuous monitoring.

1.2. BIM model – Annotations

Building Information Modeling (BIM) is a digital methodology that enables the creation and management of a building’s data throughout its lifecycle. It integrates geometric, spatial, semantic, and temporal information into a single digital model, facilitating collaboration among architects, engineers, contractors, and facility managers. BIM enhances project efficiency, reduces errors, and supports informed decision-making from design to demolition.

However, the true value of BIM is unlocked when it evolves from a static 3D model to a data-rich, interoperable digital twin. This requires enriching the model with metadata, sensor data, and contextual information—paving the way toward BIM 4.0, which aligns with the principles of Construction 4.0 and Industry 4.0 [1].

1.2.1. BIM Maturity Levels: From CAD to BIM 4.0

The evolution of BIM is commonly described through maturity levels, which reflect the degree of collaboration, data integration, and digitalization in a project. A comparative overview is presented in Table 1.

Table 1. BIM maturity levels.

Level	Description	Key Features
Level 0	No collaboration; 2D CAD only	Paper-based drawings, no data sharing
Level 1	Partial collaboration; 2D + 3D CAD	Common Data Environment (CDE), basic standards, limited model sharing
Level 2	Federated BIM; managed collaboration	Separate 3D models shared via IFC/COBie, coordinated workflows

Level	Description	Key Features
Level 3	Integrated BIM; full collaboration	Single shared model, real-time updates, cloud-based platforms
BIM 4.0	Smart, connected BIM with enriched data and automation	IoT, AI, semantic enrichment, lifecycle integration, predictive analytics

BIM 4.0 is not just a level of maturity—it represents a paradigm shift where BIM becomes a dynamic, intelligent system integrated with real-time data and advanced analytics [2].

1.2.2. The Need for Data Governance and Enrichment in BIM 4.0

To reach BIM 4.0, models must be semantically enriched with contextual data—such as sensor readings, image metadata, and operational parameters. This transforms BIM from a geometric model into a data product that supports use cases like:

- Predictive maintenance
- Energy optimization
- Safety monitoring
- Asset lifecycle management

This enrichment is achieved through ETL pipelines that automate the extraction, transformation, and integration of multimodal data (e.g., images, point clouds, IoT feeds) into the BIM environment.

1.3. Data Governance methodology

Data Governance is defined by the DAMA guide to the Data Management Body of Knowledge [3] as the exercise of authority and control (planning, implementation, monitoring, and enforcement) over the management of data assets.

It usually is also defined as a collection of roles, policies, workflows, standards, and metrics, that ensure efficient data usage and security and enables a company to reach its business objectives. It involves creating data roles and assigning permissions, designing workflows to verify information updates, ensuring data is safe from security risks, etc. [4]

Different governance frameworks and strategies include similar steps and objectives, considering the main goal to enable organizations to manage data as assets, those organizations should deploy strategies for sustainability and be able to measure the financial impact of these changes.

Best practices suggest beginning in a controlled high impact area and expand from there over time, involving at the same time, important stakeholders inside the company at the

very beginning. It is crucial to continuously adapt the system and automate tasks, when possible, to make it more efficient.

In fact, the necessity of developing a data Governance strategy comes from the importance that increasingly the companies give to their data, and, therefore, the aspects that should be covered by them. Among those aspects, there is a list of tasks that help to manage the data within the context of Governance:

- Secure your data
- Ensure compliance with regulations and data privacy laws
- Improve the data quality
- Avoid inconsistent data silos
- Improve trust in the data
- Better decision making
- Improve efficiency

Figure 1 and Figure 2, as examples, depict standard phases, commonly named in most of the methods:

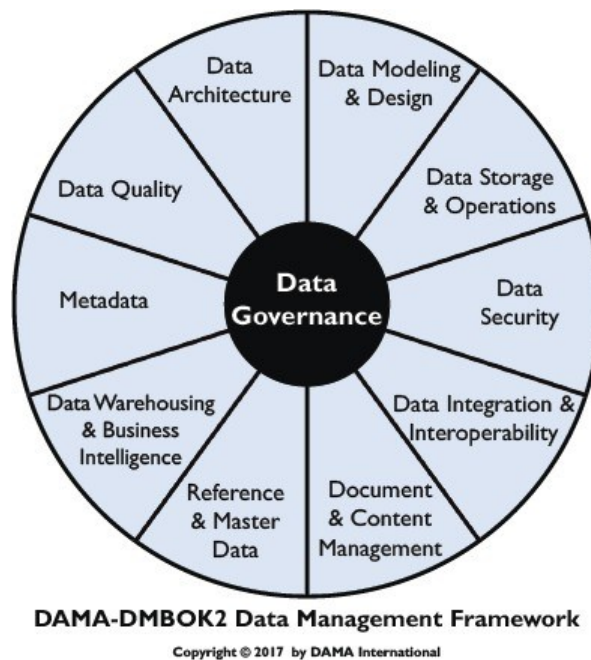


Figure 1. The DAMA DMBOK wheel. DAMA International.

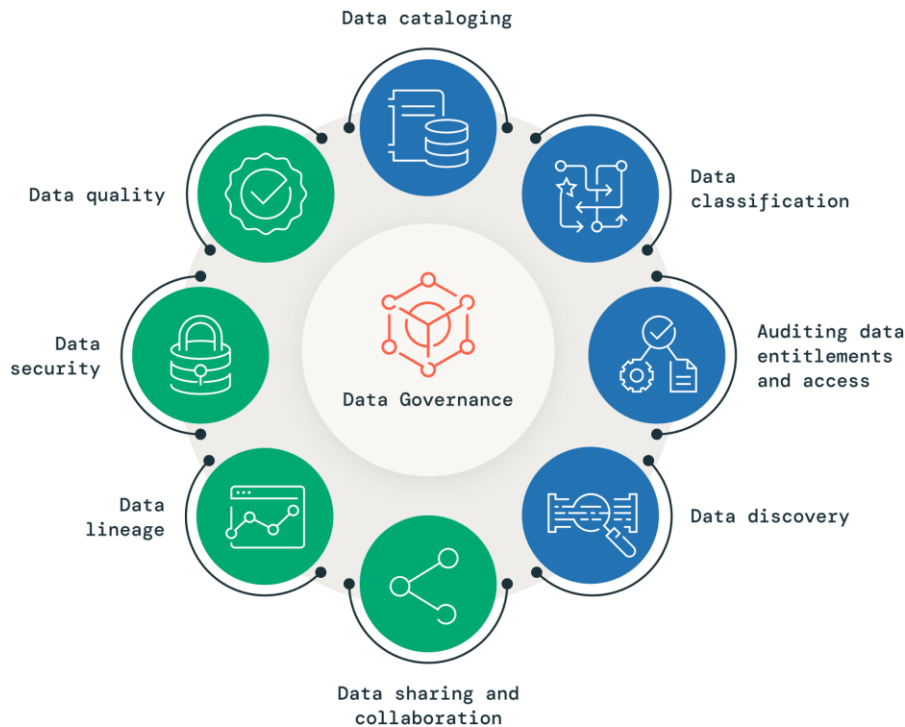


Figure 2. Key elements of Data Governance. Databricks.

The descriptions of the different elements are interconnected and share important activities that allow the organizations improving the management of their data. The following table is a summary of those activities, related to the key elements, aggregated in a general view.

Table 2. Data Governance key elements and summary description of related activities.

Data Architecture, Modelling and Design /Data discovery	Defining and managing Enterprise data models, tool standards, and system naming conventions, standard domains, and standard abbreviations. Making data easily discoverable for analytics, AI or ML use cases. Preventing data duplication.
Data Storage and Operations	Definition and use of Tool standards, standards for database recovery within the context of database performance, data retention, and external data acquisition.
Data Security/ Auditing data entitlements and access/ Data lineage/ Data sharing and collaboration	Data access security standards, monitoring and audit procedures, storage security standards, and training requirements. Data access management for data security and governance, including access controls. Comply with GDPR and CCPA for private/personal data, managing data's lifecycle providing trusted sources for audit reports. Securely and controlled data collaboration, ensuring that data privacy regulations. Data marketplaces/Data spaces for sharing data.
Data Integration & Interoperability	Selecting and implementing standard methods.
Documents and Content	Content management standards and procedures.
Reference and Master Data	Reference Data Management control procedures.
Metadata	Metadata integration procedures and usage.

Data Quality	Data quality rules, standard measurement methodologies, data remediation standards and procedures. Ensuring high data quality for accurate analytics. Evaluation of key data quality attributes (accuracy, completeness, compliance with business data-quality rules).
Big Data and Data Science/Warehousing	Data source identification, authority, acquisition, system of record, sharing and refreshing. Managing data storage and standards for Big Data handling.
Data classification	Organizing and categorizing data based on its sensitivity, value and criticality.
Data cataloging	Enabling knowledge of the data that exists within an organization. Providing a centralized metadata repository program can help organizations improve their data management, enhance collaboration, reduce redundancy and ensure proper access controls and audit information retrieval.

As another source of Data Governance definitions, Gartner contemplates [5] that the organizations should consider some key elements to ensure the effective management of the data, and considers this list as the most important one:

1. Data strategy
2. Data ownership
3. Data stewardship
4. Data policies and standards
5. Data quality management
6. Data security and privacy
7. Data life cycle management
8. Data tools
9. Compliance and regulatory requirements

Therefore, in summary, data governance is an organizational structure that supports the management of business data. Organizations must have an appropriate big data environment for storage and access and design a data architecture to govern that source data making it available to the entire company. The activities are commonly depicted in most methods.

In the context of the digital transformation of the built environment, **data governance** is a foundational element for ensuring the quality, interoperability, and semantic richness of **BIM 4.0** models. And the need to properly define this governance methodology is the proposal of this document within the framework of the DISCOVER project.

A BIM 4.0 model not only represents the geometry of an asset, but also integrates contextual, operational, and analytical information that enables process automation, informed decision-making, and integration with intelligent systems. To achieve this, a

structured data management methodology based on ETL (Extract, Transform, Load) processes and supported by **open-source technologies** is essential. This ensures traceability, standardization, and progressive enrichment of data from its origin.

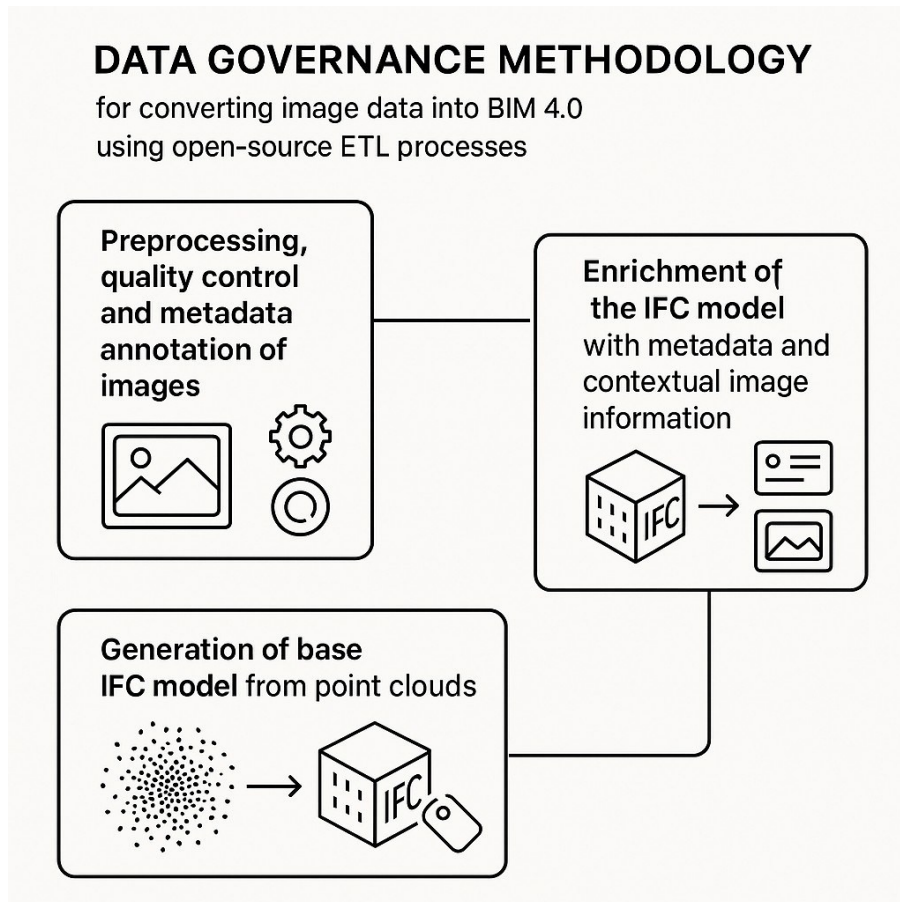


Figure 3. DISCOVER Data Governance main steps.

The methodology is structured into three key phases (Figure3):

1. **Preprocessing, Quality Control, and Metadata Annotation of Images:**

This initial phase is critical to ensure that the images used (photographs, orthophotos, thermal captures, etc.) are technically valid, free from noise or distortion, and enriched with structured metadata (e.g., geolocation, timestamp, sensor type, capture conditions). Semantic annotation—such as identifying construction elements, defects, or materials—enables automated integration with BIM models.

2. **Generation of the Base IFC Model from Point Clouds:**

The **IFC (Industry Foundation Classes)** format is an open, neutral standard developed by building SMART to represent BIM models in an interoperable way. In this phase, point clouds obtained via laser scanning or photogrammetry are processed to generate a structured, geometric representation of the environment. This base IFC model contains spatial and topological data but lacks rich semantic

information. Tools, such as **CloudCompare**, **Open3D**, **BlenderBIM**, and **IfcOpenShell**, are used to convert point clouds into IFC-compatible models.

3. Enrichment of the IFC Model with Metadata and Contextual Image Information:

In the final phase, metadata and contextual information extracted from images are integrated into the IFC model, associating visual, technical, and operational data with each model component. This enables the transition to a **BIM 4.0** environment, where each object is not only geometrically defined but also semantically and operationally enriched. This enrichment can be automated using **Python** scripts and use cases databases for linking external data sources.

1.4. Phases, roles and responsibilities

The different phases or elements described as part of the Data Governance Methodology should have a person or team associated with promoting the optimal achievement of their objectives. In general, the main roles that the methodologies consider are presented in Figure 4.

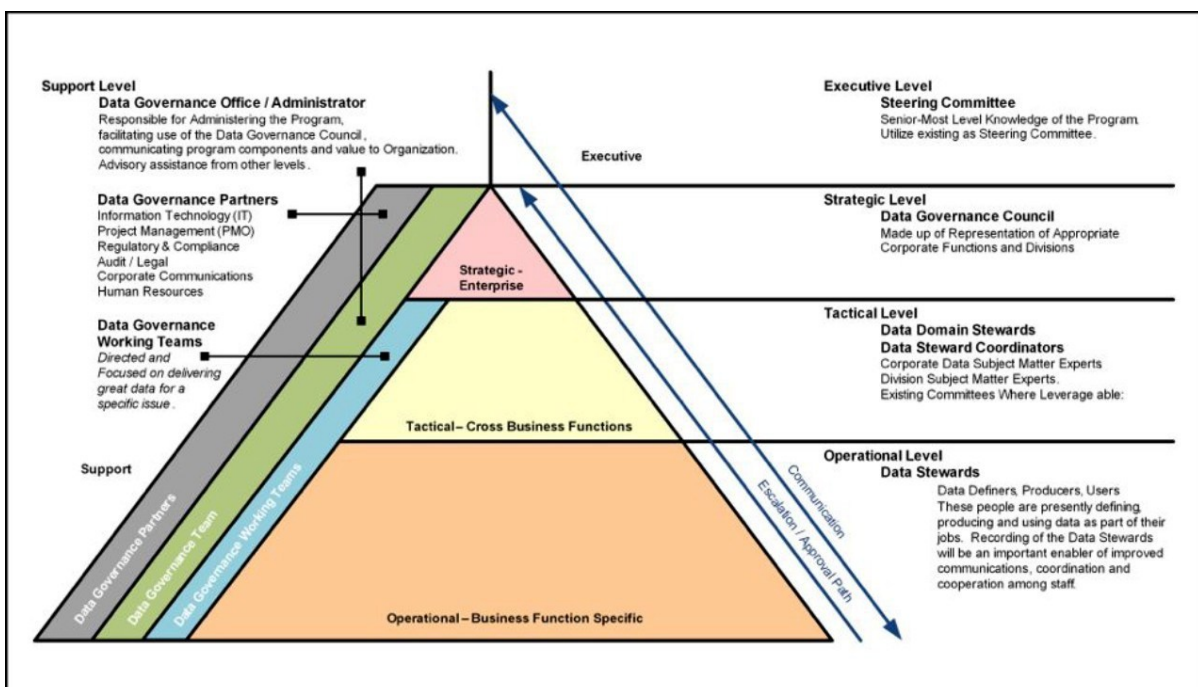


Figure 4. Governance Roles and Responsibilities. Robert S. Seiner, KIK Consulting.

The scope of the Governance workflow for the DISCOVER project is not as formal and extensive as those methodologies describe. Since the implementation of the data life cycle will be managed through a pipeline as an ETL process (Figure 5), the roles and responsibilities will be not such a defined structure. Each partner, as data owner, would participate as Executives and Tactical actors, whereas the coordinator of the technical implementation of the pipeline will act as Operational actor of the whole deployment.

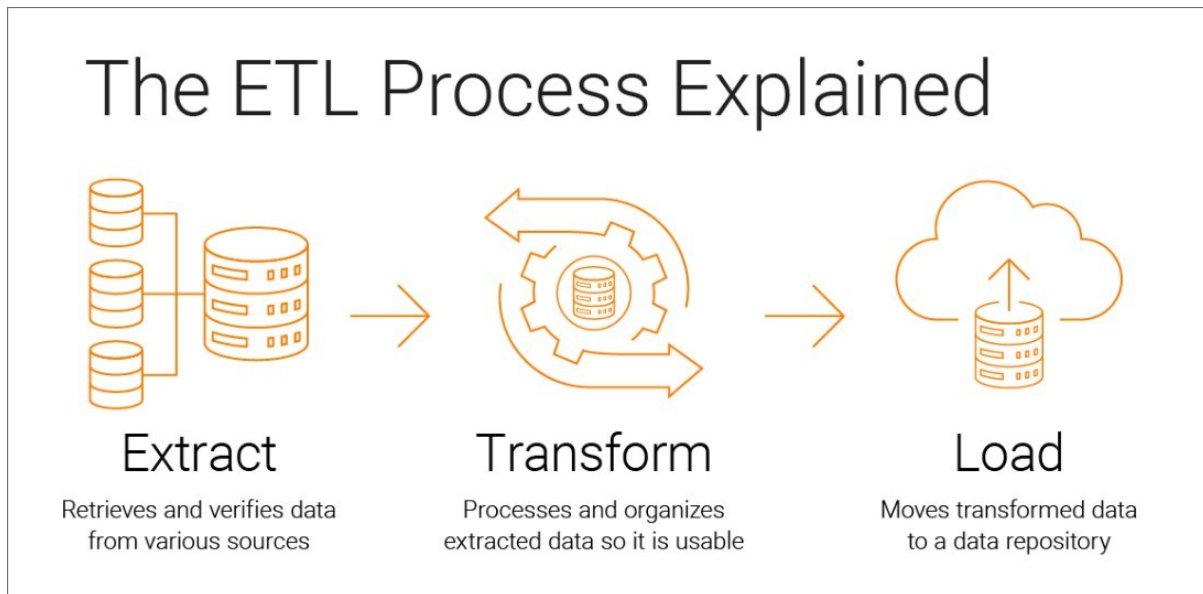


Figure 5. ETL process. Informatica.com. [6]

The Extract, Transform and Load process for managing the data follows a methodology consisting of three stages:

- **Extract:** For collecting raw data from different data sources, whom types could be Structured (SQL databases, ERPs, CRMs), Semi-structured (JSON, XML) and Unstructured (e-mails, web pages, files).
- **Transform:** Where the data could be cleaned, aggregated, or enriched following the rules defined by the organization.
- **Load:** Transferring the data after transformations into a storage system.

Embedding governance practices into ETL processes is a good strategy to ensure the accuracy of the data. General activities in data governance could be related to some functions in an ETL. How do the ETL pipelines cover the Data Governance Steps? Figure 6 illustrates the activities configuring in a Data Governance strategy and the corresponding actions to be performed by the ETL pipeline.

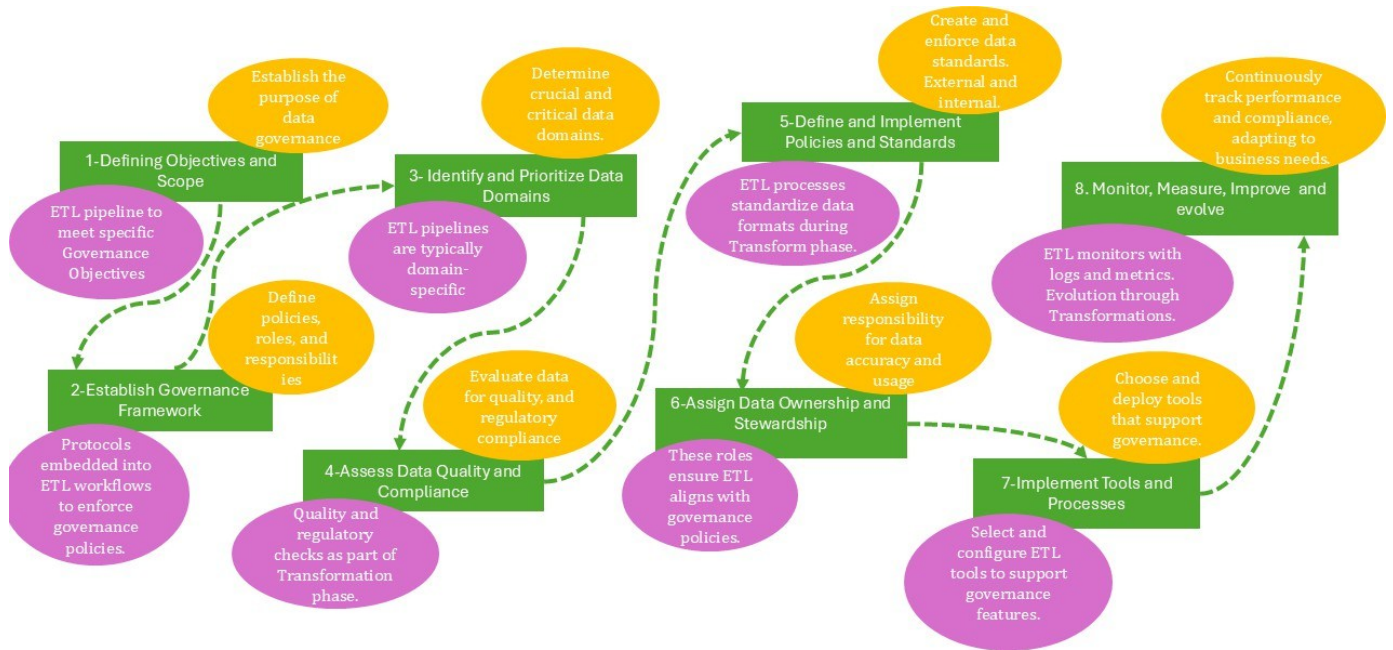


Figure 6. ETL and Data Governance processes.

In summary, ETLs support the Data Governance methodology for managing the data life cycle, and they must be built to meet the more important benefits that the organizations consider. For example:

- Tracking the origin and transformation of datasets.
- Developing automated checks for data quality as part of the Transform phase.
- Implementing access controls to the data.
- Including logs for compliance and troubleshooting.
- Involving data stewards in ETL design and validation.

The Figure 7 illustrates the main three phases and the corresponding ETL pipelines:

1. ETL1 - Metadata & Image Annotation

- Emphasizes metadata extraction and personalized transformation.
- Image metadata annotation (e.g., CSV format with fields like Id, Location, Timestamp, Device).

2. ETL2 - IFC Generation from Point Clouds

- By means of point cloud input realizes the IFC generation.
- Important characteristics are the point of cloud density, format, and typology mapping.

3. ETL3 - IFC Enrichment

- Image metadata is joined with IFC elements.
- Final output expected is a semantically enriched **BIM 4.0** model.

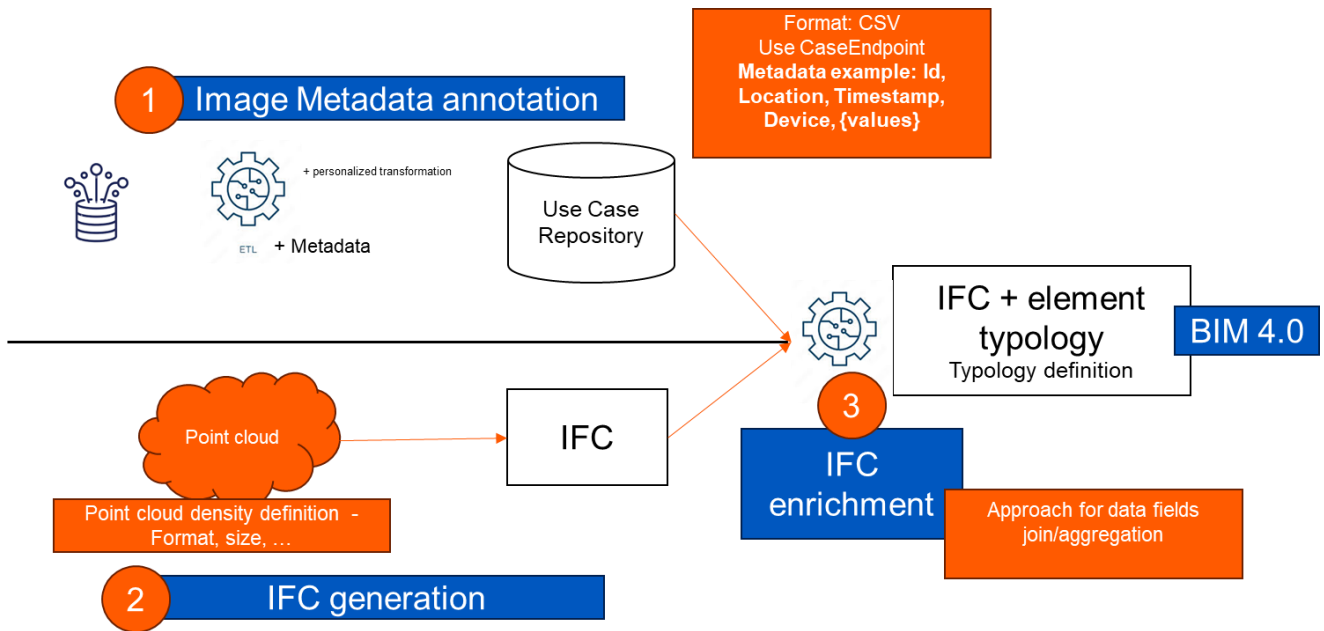


Figure 7. Main phases of DISCOVER BIM 4.0 generation approach.

The following table provides a structured overview of the three ETL phases that underpin the data governance methodology implemented in the DISCOVER project. Each ETL process plays a critical role in the transformation and integration of data—from image metadata extraction to the generation and semantic enrichment of IFC models—ensuring consistency, traceability, and interoperability across all use cases. This breakdown serves as a key reference for monitoring responsibilities, data sources, and stakeholder involvement, supporting transparent and efficient project execution.

Table 3. Summary of ETL phases in DISCOVER.

Phase	Objective	Description	Execution Responsibility	Data Providers	Participants
ETL1	Metadata extraction and image annotation	Extraction of relevant image metadata and annotation in structured format (CSV), including fields like Id, location, timestamp, etc.	Tecnalía with the DISCOVER consortium partners	Image providers from the different use cases	Tecnalía + DISCOVER consortium partners
ETL2	IFC model generation from point clouds	Processing of point clouds captured within the project to generate IFC models, considering	DTT	Point clouds captured within the DISCOVER project framework	DTT

		density, format, and typology mapping.			
ETL3	Semantic enrichment of the IFC model for BIM 4.0	Integration of image metadata with IFC elements to produce a semantically enriched BIM 4.0 model usable across all project use cases.	Collaboration between DTT and Tecnia	Data from ETL1 (metadata) and ETL2 (IFC models)	DTT + Tecnia

1.5. Phase 1: Metadata generation ETL

ETL1 focuses on the extraction and structured annotation of image metadata, serving as the foundational step in the DISCOVER project’s data processing pipeline. This phase involves collecting metadata such as image ID, location, timestamp, and device information, typically formatted in CSV files. The process is led by Tecnia in collaboration with the DISCOVER consortium partners, who act as image providers across the various use cases. The goal is to ensure that all visual data is enriched with standardized, high-quality metadata, enabling downstream interoperability and traceability. ETL1 plays a critical role in aligning data inputs with the project’s governance methodology, ensuring consistency and readiness for integration in later stages of the BIM 4.0 model development.

In the following sections, we will delve deeper into the specific artifacts, data sources, and transformation processes involved in this phase. This detailed breakdown will provide greater insight into how ETL1 supports the overall data governance strategy of the DISCOVER project.

1.5.1. Main artifacts (Information to be gathered from partners in an Excel sheet)

The data that will be managed along the ETL implementation are provided by the different partners in the project. At this moment of the project, these datasets are being described from TU Delft:

- DS_2_1_TUD_XRF
- DS_2_2_TUD_RGB_Invasive
- DS_2_3_TUD_NIR
- DS_2_4_TUD_RGB_NonInvasive

These datasets are provided by UPC:

- Point cloud data for the geometry of the building

- Building geometry (CDEI-UPC)
- "Semantic points"
- Robot trajectories (if needed for visualization)

The following sections will describe the preliminary aspects of these datasets and the transformations to be applied to them to obtain the final datasets, considered as Final Data Products.

1.5.2. Data sources – origin of data

The information about how the data will be gathered and entered the pipeline is not available at this point of the project. Nevertheless, the partners involved provided several descriptions of the data and how they should be managed to depict in a conceptual manner, the principal artifacts in the ETL pipeline. Further deliverables will include references to the evolution of the ETL design and deployment.

- DS_2_1_TUD_XRF:

Structured data with 21 pre-processed features is ready for machine learning training. Each sample contains the concentration of key elements that make up the material, such as Mg, Al, P, S, K, Ti, and others. An AI model will be built using this dataset.

The origin is a XRF scanner, and the type of data is Structured-Tabular Data.

- DS_2_2_TUD_RGB_Invasive

An unstructured data source will be used alongside XRF to enhance the identification of drilled powder materials based on their textural information. An AI model will be developed using this dataset.

The origin is a RGB image sensor, and the type of data is un-structured image data (will require a separate processing pipeline with different transformations).

- DS_2_3_TUD_NIR

No information provided yet.

- DS_2_4_TUD_RGB_NonInvasive

No information provided yet.

- Point cloud data for the geometry of the building.

The origin is an Ouster Dome Lidar.

- Building geometry (CDEI-UPC)

The origin is Ouster Dome Lidar + camera -> RTAB-MAP.

- "Semantic points"

The origin is GPR, RGB camera, and other sensors later.

- Robot trajectories (if needed for visualization, etc)

The origin is Robot localization.

1.5.3. Transformations

TU Delft has the intention of building the AI models, so all the necessary preprocessing and transformation steps are included along with the algorithms. This means no additional cleaning, preprocessing, or transformation will be performed in previous steps and the data gathered from the initial point of the ETL (Data Source) will be "transformed". The final datasets will be ready and pre-processed for direct integration into the main project pipeline for specific AI tasks and they will require fewer transformations before being fed into any AI model.

Additionally, the type of data is an important information that could impact the metadata describing them. In this case, two types of data are considered:

1. Structured data (tabular format), which is currently collected from XRF and may include NIR data in the future.
2. Unstructured data from image datasets, which will require a separate processing pipeline with different transformations.

The type of data provided by UPC could be used as important criteria to define further transformations, not envisioned at this moment.

1. Point cloud data have been registered as (x, y, z) coordinates.
2. Building geometry consists of .ply files.
3. Semantic points will be conformed as a set of fields [(point id, element that point belongs to, material, size, coordinates, reliability). For instance: (point1, wall1, concrete, wall depth value, (x1, y1, z1), 98%)].
4. Robot trajectories, registered as coordinates in this structure (x, y, z, orientation).

The ETL pipeline will provide general transformations related to the quality of the data, such as accuracy, completeness, and compliance with business data-quality rules, if any.

1.5.4. Final Data Product - Minimum set of metadata

The final step of ETL1 involves populating a dedicated database or repository for each use case with the processed and annotated image data. This ensures that all relevant metadata is stored in a structured and accessible manner, aligned with the data governance framework defined in the DISCOVER project. Each image entry includes a minimum set of metadata fields considered essential for downstream processing and integration:

- Identifier
- Location
- Timestamp
- Capturing device or equipment

The output of this step is a comprehensive dataset consisting of all annotated images, each represented in JSON format (Figure 8). This structured collection serves as the authoritative source of image metadata for subsequent ETL phases and guarantees consistency, traceability, and semantic alignment across all use cases.

```

1  {
2    "image_id": "IMG_20250718_00123",
3    "location": {
4      "latitude": 43.3051,
5      "longitude": -2.8859,
6      "site_name": "UseCase_Area_1"
7    },
8    "timestamp": "2025-07-18T10:45:00Z",
9    "capturing_device": {
10     "device_id": "CAM_XYZ_01",
11     "device_type": "RGB Camera",
12     "provider": "PartnerOrg_A"
13   }
14 }
15

```

Figure 8: Example of Image Metadata JSON.

1.6. Phase 2: IFC Generation

ETL2 is dedicated to transforming raw point cloud data into structured IFC models, a critical step in enabling semantic interoperability within the DISCOVER project. This process is led by DTT, using point clouds captured specifically within the framework of the project. The transformation involves analysing key characteristics of the point clouds, such as:

- Density: the level of detail captured in the spatial data;
- Format: compatibility with processing tools and IFC conversion workflows;
- Typology mapping: classification of spatial elements (e.g., walls, floors, equipment) to align with IFC schema.

The output of ETL2 is a set of IFC models that represent the physical environments of each use case in a standardized, machine-readable format. These models serve as the structural backbone for the subsequent enrichment phase (ETL3), where semantic data from images and metadata will be integrated. ETL2 ensures that spatial data is accurately and

consistently modelled, forming a reliable foundation for the BIM 4.0 vision of the DISCOVER project.

1.6.1. Point Cloud Data Preparation Protocol for BIM Development

In T3.4 the partners will develop BIM models in IFC format from the point clouds collected in WP1. The collection of high-quality point cloud data is essential for the generation of accurate BIM models. This section outlines the data preparation protocol for point cloud acquisition to ensure that the data provided by WP1 are suitable for reliable and efficient BIM reconstruction at Level of Development (LOD) 200. LOD 200 models include generalised geometry that approximates the size, shape, and location of building components.

Following cleaning and filtering of the raw point cloud data, the partners will apply semantic segmentation and object detection using deep learning algorithms to classify the points in structural building elements (walls, floor, columns, doors, etc.). To enable the successful labelling, prior training of the deep learning algorithms on annotated datasets is required. The partners intend to utilise the Stanford 3D Indoor Scene Dataset (S3DIS) [7], which contains six large-scale indoor areas with 271 rooms, for training deep learning algorithms. Each point in the S3DIS is annotated and consists of three-dimensional spatial coordinates (XYZ) and Red-Green-Blue (RGB) colour information. Therefore, it would be expedient if the point clouds collected in WP1 using 3D Light Detection and Ranging (Lidar) scanners and RGB-D cameras capture these six attributes (X, Y, Z, and RGB). If only spatial coordinates are collected, the approach remains feasible but may decrease segmentation accuracy.

The collected point clouds must be georeferenced or, if applicable, aligned to a local building coordinate system. Accurate spatial referencing is essential for the registration of different point clouds and for the integration of multi-room or multi-floor datasets into a coherent BIM model. Adjacent scans should have a minimum overlap of 30% to enable precise registration and alignment. To combine individual point clouds, it is helpful to use reference markers during the scanning process. All data should be accompanied by comprehensive metadata, including information about the acquisition conditions, sensor specifications, and coordinate system definitions.

To ensure interoperability across different processing platforms and modelling tools, the recommended file format is a plain text file (txt) or the Polygon File Format (ply). The S3DIS is stored in a .txt file. Therefore, the same format would be compatible and easy to use. ply files are commonly used and can store both geometric (XYZ) and colour (RGB) data. While there are only common conventions for storing txt files from point clouds, ply files can be saved either as human-readable American Standard Code for Information Interchange (ASCII) files or in binary format, which is more efficient for computational processing. Thus, both formats are well-suited for deep learning applications.

Completeness of the dataset is a key requirement. Point clouds must comprehensively capture all accessible structural features of the built environment. Practical challenges,

such as the presence of furniture or limited scanner lines of sight, may restrict full coverage. To mitigate occlusions, scans should be captured from multiple angles. Insufficient coverage of critical surfaces—such as wall junctions—may lead to incomplete semantic segmentation results and require more manual intervention, thereby, reducing the effectiveness of the intended (semi-)automated BIM reconstruction pipeline. Therefore, a lower coverage of structural elements is aimed for, but considering these real-world circumstances is an important aspect for testing our approach at the indoor demonstration sites.

To ensure reliable results, point clouds must have at least a millimetre level of accuracy for developing architectural models [8]. Therefore, the positional accuracy of the captured points should be within ± 10 millimetres, ideally around ± 5 millimetres. This tolerance is necessary to ensure that the derived BIM models accurately reflect the as-built conditions, particularly for architectural and structural elements. In applications where finer detail is required, especially for smaller components or specialised analyses like pipes, a tighter tolerance is recommended.

Point spacing and density are also important factors when developing BIM models from point clouds, as they directly affect the accuracy and completeness of the resulting model. For developing BIM models at LOD 200, an average point spacing between 5 and 10 millimetres is expedient to avoid gaps that could hinder surface reconstruction and enable precise feature extraction during modelling. This corresponds to a point density ranging from 10,000 to 40,000 points per square meter. Areas with complex geometry or fine details may require higher densities to capture essential features accurately.

In summary, the preparation of high-quality, accurate, and semantically enriched point cloud data is fundamental to the success of developing BIM models from point clouds. While the mentioned requirements serve as reference values, more detailed insights due to data collection (higher positional accuracy, point densities, etc.) will ensure the efficient development of accurate BIM models, but require more storage capacity.

1.7. Phase 3: IFC Enrichment

ETL3 represents the final phase of the DISCOVER project's data governance and transformation pipeline, where the previously generated IFC models (ETL2) are semantically enriched using the metadata extracted and annotated in ETL1. This phase is carried out collaboratively by DTT and Tecnalia, with the goal of producing a unified and intelligent BIM 4.0 model that can be reused across all project use cases.

The enrichment process involves joining the image metadata—stored in structured JSON format—with the corresponding elements in the IFC model. Each metadata entry includes key fields such as:

- **Identifier**
- **Location**

- **Timestamp**
- **Capturing device**

To perform this integration effectively, the best approach for joining the data is to use a spatial-temporal matching strategy, which combines:

1. **Spatial alignment:** Matching the location coordinates of the image metadata with the spatial geometry of IFC elements;
2. **Temporal correlation:** Using timestamps to associate images with construction phases or inspection events;
3. **Device context:** Linking the capturing device to specific viewpoints or scanning sessions.

This multi-dimensional join ensures that each IFC element is enriched with contextual information, enabling advanced querying, visualization, and decision-making capabilities within the BIM 4.0 environment. The result is a semantically rich model that integrates both geometric and descriptive data, fully aligned with the DISCOVER project's data governance methodology.

1.8. Reference Architecture: Implementation design

The technological architecture designed to support the ETL processes in the DISCOVER project serves as a foundational template that enables the implementation of robust data governance across diverse use cases. This architecture provides a modular and scalable framework that ensures the consistent execution of data extraction, transformation, and loading workflows, while maintaining alignment with the project's governance principles.

Rather than being a rigid, one-size-fits-all solution, the architecture is conceived as a flexible blueprint—a reference model that can be adapted, implemented, and deployed according to the specific needs and infrastructure of each use case. Whether the data originates from image capture, point cloud scanning, or semantic modelling, the architecture ensures that all components are interoperable, traceable, and ready for integration into the BIM 4.0 ecosystem. This adaptability is key to supporting the multimodal and collaborative nature of the DISCOVER project, while maintaining a unified approach to data quality, security, and lifecycle management.

1.9. System Requirements

In modern data-driven environments, Extract, Transform, Load (ETL) architectures play a critical role in ensuring that data flows reliably from diverse sources to centralized repositories such as data warehouses or data lakes. However, the increasing demand for high-quality, trustworthy, and well-governed data has raised the bar for what an ETL architecture must deliver.

This section outlines the key functional and non-functional requirements that a robust ETL architecture should fulfil to meet enterprise standards for **data quality, data governance, metadata management, and transformation logic**.

Functional requirements:

- **Extraction:** Support for various data sources, like APIs, SQL or No-SQL.
- **Transformation:** Apply validation and business rules with full lineage tracking.
- **Data Storage & Sharing:** Reliably save processed data in various destinations and enable easy sharing across teams and systems with error handling.
- **Metadata Management:** Capture and version both technical and business metadata.
- **Data Quality:** Enforce customizable rules for accuracy, completeness, etc., with alerts and quarantine.
- **Governance:** Role-based access control, auditing, data cataloguing, and integration with governance tools.

Non-functional requirements:

- **Scalability:** Ability to handle large volumes with distributed processing.
- **Availability & Resilience:** High availability, fault tolerance, and recovery mechanisms.
- **Observability:** Real-time monitoring, centralized logging, alerting.
- **Maintainability:** Modular pipeline design, low-code configuration, and automated testing.

Documentation and Usability requirements:

- Automatically generate clear and up-to-date documentation for ETL pipelines and metadata.
- Provide an easy-to-use interface for users to explore data definitions and pipeline details.
- Maintain a shared data dictionary accessible to both technical and business users.
- Support version control for all documentation and metadata changes.

1.10. Functionalities and components

The system is composed of two main components: Edge and Cloud (Figure 9). The Edge is responsible for collecting and processing data, which is then transmitted to the cloud. This data can be in several formats:

- Raw format;
- Transformed format, where the Edge performs transformations and sends tabular data to the cloud;
- Sensor-specific file format.

Edge can either actively send the data to the cloud or store it in a database (SQL or non-SQL, depending on the data format). The cloud retrieves the data from the database and ingests it. The architecture is illustrated in the figure, which highlights the different methods the cloud uses to process data.

Two essential processes, Extract and Load (EL), are always performed. Data transformation is only necessary when Edge hasn't transformed the data or when it's not in a sensor-specific file format. To ensure optimal understanding and management of stored data, metadata from all data types is stored in a tabular format.

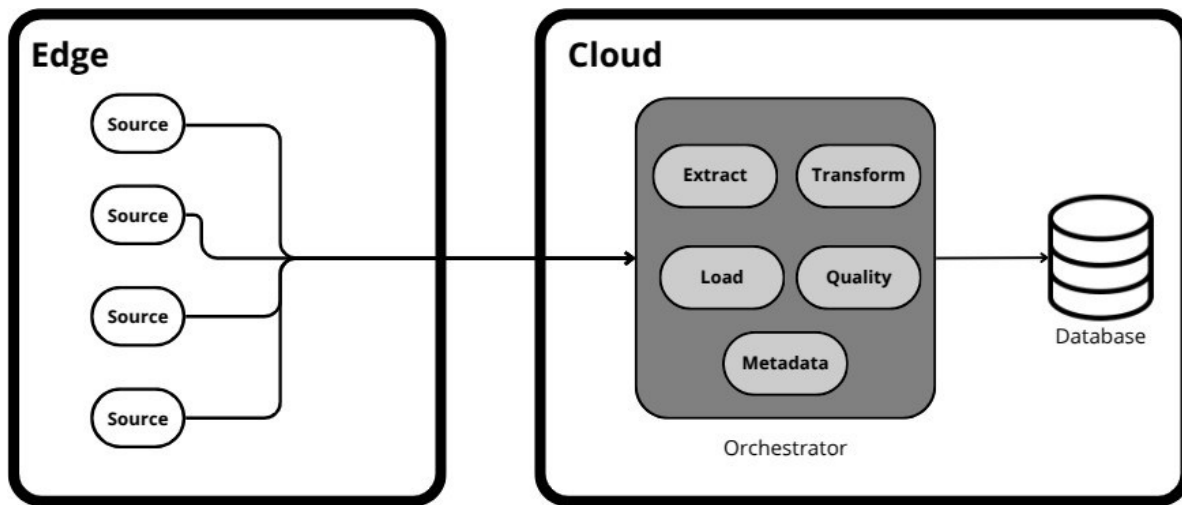


Figure 9. ETL General architecture.

1.11. Technology stack

In this section, the comparison between different technologies of the cloud is going to be performed to decide which is the best for the project. The first step is to identify the functionalities that these tools must fulfil:

1. API for the communication with the Edge.
2. Orchestrator to perform ETL operations.
3. No-SQL database for sensor-specific file format.
4. SQL database for tabular data and metadata.

To develop an API, currently the best option is **Fastapi**, as it has a high performance for real-time applications, can handle multiple requests concurrently, has tight integration with Python ecosystem and it is simple, highly customizable and consistent.

Several options are available for orchestrators, but Apache Airflow, Mage and Luigi are the most relevant choices for this specific use case. The table below provides a detailed comparison of their key features and capabilities.

Table 4. Comparison of orchestrator software options.

Feature	Apache Airflow	Mage	Luigi
UI/Web Interfaces	Very complete	Modern and intuitive	Limited
Time-based Orchestration	Yes (cron, DAGs)	Yes (intuitive scheduling)	Yes
Easy to use	Medium (learning curve)	High (user friendly)	Medium to Low
Extensibility	Very High (plugins, operators)	High (modular)	Medium
Scalability	Very High (Celery)	High	Limited
Logging	Very Good	Good	Basic

Based on the comparison, considering **Apache Airflow**'s exceptional logging capabilities, scalability, and extensibility, it emerges as the most suitable choice for our project specific use case.

For selecting a No-SQL database, several options are available, including key-value, document-based, columnar, and time-series databases. However, for this specific use case, document-based databases are the most suitable choice due to their flexibility in handling diverse data types. **MongoDB** stood out as the top choice, providing a robust scalable solution.

For selecting SQL database, as shown in the comparison table above, **PostgreSQL** clearly stands out among the top SQL-databases. While MySQL and MariaDB are both power and widely adopted, PostgreSQL excels in areas critical for advanced applications: standards compliance, data integrity, extensibility and performance. Its performance with complex queries, advanced indexing options, and strong adherence to the SQL standard make it deal for high-load environments and data-heavy applications.

Table 5. Comparison of open source relational databases.

Feature	PostgreSQL	MySQL	MariaDB
SQL standard compliance	Very high	Medium	Medium-high
Extensibility	Very high	Medium	High
Storage engine	One highly optimized main engine	Multiple engines	Multiple engines
Extensibility	Very High (plugins, operators)	High (modular)	Medium
Community	Very active	Very active	Very active

For sensors with sensor-specific file formats, an FTPS server is required to receive and store their files, which the orchestrator can then process accordingly. To achieve this, there have been identified three viable options: FilleZilla, vsftpd and ProFTPD. A comparison of these alternatives is presented in the following table.

Table 6. Comparison of FTP client and server software.

Feature	FilleZilla	vsftpd	ProFTPD
FTPS support	Yes	Yes	Yes
Graphical interface	Yes	No	No
SSL/TLS configuration	Intuitive via UI	Manual via config files	Manual via config files
Support and community	Active and strong	Active	Active
Community	Very active	Very active	Very active
Deployment	Easy	Easy	Medium

Based on the table, both FilleZilla and vsftpd emerge as top contenders. However, considering the ease of use provided by its graphical interface and the support of a strong community, **FilleZilla** is the best option for this system.

The systems architecture is visually represented in the diagram below (Figure 10), which illustrates the cloud component and highlights how the selected technologies, FastAPI, Apache Airflow, MongoDB, PostgreSQL and FilleZilla work together to form a cohesive and scalable infrastructure.

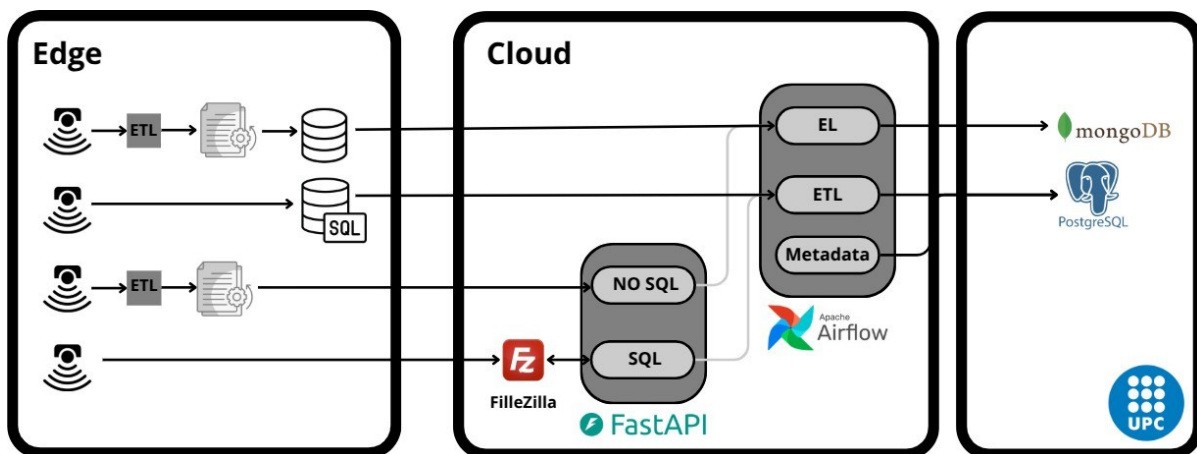


Figure 10. Cloud components of the ETL system.

1.12. Infrastructure

The design of the general schema of the ETL is depicted in Figure 11. Each of the Use Cases starts with a deployment of their pipeline, meeting their specific requirements and storing their data in the repositories, ready to be analysed and transformed to obtain the outcome in the form of an IFC schema. Each use case is deployed independently within the infrastructure of each project partner, following a standardized yet adaptable structure. Tecnalía provides a centralized ETL Testing Environment, and a set of reusable templates managed through a shared GIT repository. These templates include predefined components and workflows to facilitate the integration and deployment of data processing pipelines across different partners and contexts. In each specific use case, raw

data is ingested and processed through local ETL workflows, where it is transformed, cleaned, and enriched with relevant metadata. This ensures that each partner can tailor the ETL process to their own infrastructure and data sources, while still maintaining compatibility with the overall framework. The processed data and metadata are stored in a local Use Case Repository managed in collaboration with UPC. These repositories act as the foundation for generating and enriching IFC models. Finally, DTT and Tecnalia will leverage the structured data from these repositories to support advanced BIM 4.0 functionalities, such as semantic enrichment, interoperability, and automated generation of standardized IFC outputs. This modular and scalable architecture promotes interoperability, reusability, and consistent data integration across a wide range of real-world scenarios.

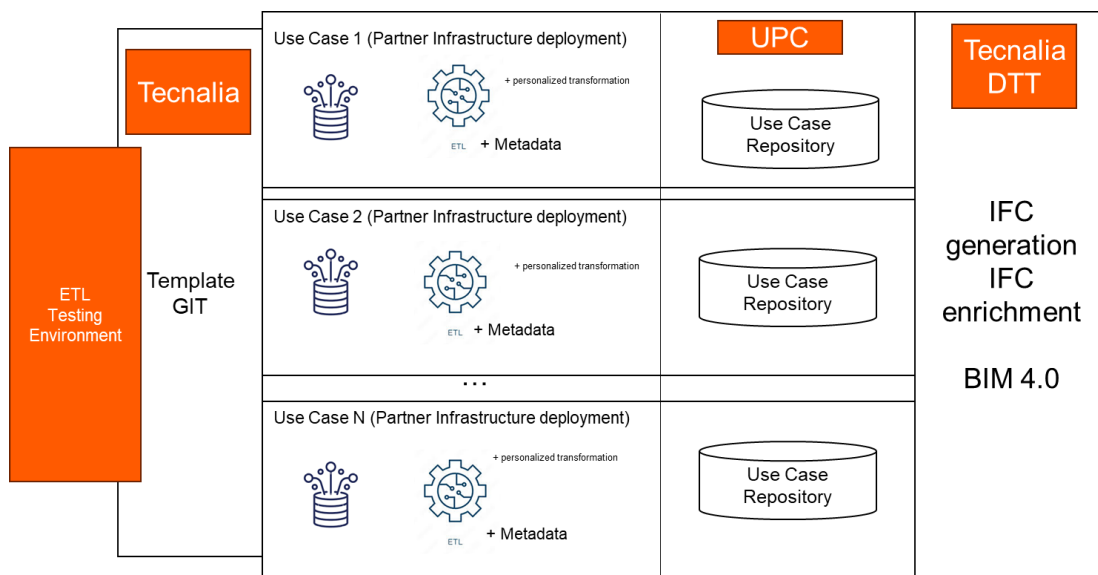


Figure 11. Conceptual Design of the Use Cases infrastructure.

1.12.1. ETL testing environment description

The ETL architecture will be deployed on an OpenStack platform, leveraging its virtualized compute resources and container orchestration for flexible workload management. Storage is provided through block and object storage, ensuring reliable data persistence and accessibility. Networking enables isolated, secure communication with load balancing and integration capabilities.

This project involves the deployment and configuration of four distinct virtual machine instances within an OpenStack cloud environment. The instances are named as follows:

- **discover_master_instance**
- **discover_connector_instance**
- **discover_data_instance**
- **discover_etl_instance**

A detailed view of the architecture is provided in the diagram below (Figure 12), which illustrates the instance-level configuration and highlights the key technologies used in each component.

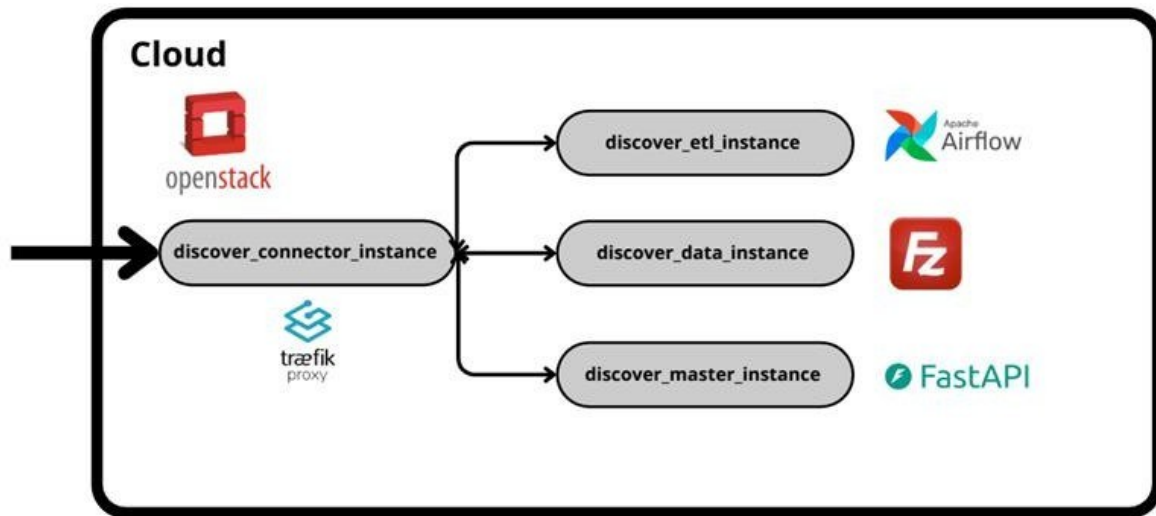


Figure 12. Instance-level configuration of the system.

Each instance serves a specialized role to ensure a modular, scalable, and maintainable system architecture.

1.12.2. Instance Descriptions

1. discover_master_instance

The **discover_master_instance** acts as the central orchestrator of the entire system. It manages coordination, scheduling, and control functions, handling requests and distributing workloads to the other instances.

Advantages:

- Centralized control simplifies management and monitoring.
- Improves coordination across different components, enhancing system reliability.

2. discover_connector_instance

The **discover_connector_instance** is responsible for interfacing with external data sources or services. It handles connectivity, data ingestion/extraction, and protocol translation to ensure seamless communication between external systems and the internal infrastructure.

Advantages:

- Decouples data source integration from processing logic.
- Makes the system more adaptable to various data sources and protocols.

- Simplifies troubleshooting and updating connectors without impacting other components.

3. discover_data_instance

The **discover_data_instance** is dedicated to storing, managing, and providing access to data and metadata. It may run database services or file storage systems optimized for efficient querying and retrieval.

Advantages:

- Centralizes data storage for consistency and security.
- Allows independent scaling of storage resources based on demand.
- Supports optimized data access patterns suited to workload requirements.

4. discover_etl_instance

The **discover_etl_instance** handles the ETL processes. It is responsible for data cleaning, transformation, enrichment, and loading data into the **discover_data_instance** for consumption by applications or analytics.

Advantages:

- Isolates ETL processing to avoid impacting other system operations.
- Facilitates modular development and maintenance of data pipelines.
- Enables efficient resource usage by tuning ETL tasks independently.

1.12.3. Advantages of Using OpenStack for Deployment

- **Flexibility:** OpenStack enables quick virtual machine instance deployment and customization to meet the requirements of each component.
- **Scalability:** The ability to scale each instance independently based on workload enhances cost-effectiveness and resource usage.
- **Isolation:** Enhancing security and fault isolation involves running distinct instances for various components. A compromise or failure in one situation does not immediately impact others.
- **Open Source and Community Support:** OpenStack is a popular open-source cloud technology that provides regular upgrades and a large community.
- **Automation:** To improve operational efficiency, OpenStack offers deployment, monitoring, and lifecycle management automation technologies.
- **Cost-effective:** By using OpenStack on private clouds or commodity hardware, reliance on pricey proprietary cloud services is decreased.

Conclusions

This deliverable has presented a comprehensive data governance methodology tailored to the needs of the DISCOVER project, with a particular focus on the design and structuring of ETL pipelines to support the generation of semantically enriched BIM 4.0 models. The proposed governance framework ensures traceability, interoperability, and quality control across the data lifecycle, while the accompanying technological architecture provides a robust and modular foundation for implementation.

It is important to note that the architecture described herein is intended as a **reference template**—a flexible and adaptable blueprint that must be customized and deployed according to the specific requirements and infrastructure of each use case. As such, this document does not prescribe a fixed implementation, but rather outlines best practices, system requirements, and technological recommendations that serve as a starting point for further development. The actual deployment and operationalization of the ETL pipelines will be the responsibility of each use case team and may involve adjustments or extensions to the methodology to ensure practical applicability and usability in real-world conditions.

Finally, while the document outlines the expected data sources and transformation processes, it is acknowledged that **many of the datasets referenced have not yet been generated** at the time of writing. Further work is required to consolidate and validate the data acquisition pipelines, particularly in relation to the origin and structure of the input datasets. These future steps will be essential to fully realize the potential of the proposed data governance strategy and to produce high-quality, annotated datasets that can drive the DISCOVER project's objectives forward.

References

- [1] Sajjad Ahmadiania, "Construction 4.0 in BIM," *SATABIM | Design & Build*, Aug. 23, 2023. <https://satabim.com/construction4-in-bim/> (accessed Aug. 31, 2025).
- [2] Y. Chen, D. Huang, Z. Liu, M. Osmani, and P. Demian, "Construction 4.0, Industry 4.0, and Building Information Modeling (BIM) for Sustainable Building Development within the Smart City," *Sustainability*, vol. 14, no. 16, p. 10028, Aug. 2022, doi: <https://doi.org/10.3390/su141610028>.
- [3] "DAMA® Data Management Body of Knowledge (DAMA-DMBOK®)," *DAMA International®*, May 27, 2025. <https://dama.org/learning-resources/dama-data-management-body-of-knowledge-dmbok/>
- [4] "Data Ladder," *Data Ladder*, Jun. 21, 2021. <https://dataladder.com/>
- [5] "Understand Data Governance Trends & Strategies | Gartner," *Gartner*, 2024. <https://www.gartner.com.au/en/data-analytics/topics/data-governance> (accessed Aug. 31, 2025).
- [6] "What is a Data Pipeline? Definition and Best Practices," *Informatica*. <https://www.informatica.com/resources/articles/data-pipeline.html>
- [7] I. Armeni *et al.*, "3D Semantic Parsing of Large-Scale Indoor Spaces," *IEEE Xplore*, Jun. 01, 2016. <https://ieeexplore.ieee.org/document/7780539>
- [8] C. Wang, C. Wen, Y. Dai, S. Yu, and M. Liu, "Urban 3D modeling with mobile laser scanning: a review," *Virtual Reality & Intelligent Hardware*, vol. 2, no. 3, pp. 175–212, Jun. 2020, doi: <https://doi.org/10.1016/j.vrih.2020.05.003>.