Point Cloud Data Management for Cross-Domain Applications

Kwasi Nyarko Poku-Agyemang[1], Balthasar Teuscher[1], Martin Werner[1]

[1] Technical University of Munich, Germany; TUM School of Engineering and Design, Department of Aerospace and Geodesy, Professorship of Big Geospatial Data Management.
(kwasi.poku@tum.de, balthasar.teuscher@tum.de, martin.werner@tum.de)

Keywords: Point Cloud, Data Model, Big Data Analysis, Data Management.

Abstract

Point clouds have proven over the years to be a suitable spatial representation of scenes and objects at varying scales and levels of complexity, making them widely used across several scientific domains and applications. Advancements in sensor technology, computer vision, and data science have produced high-resolution point clouds and advanced analytical approaches, leading to broader adoption for spatial information extraction to support decision making. However, traditional point cloud management systems for organizing and distributing data throughout the point cloud lifecycle often create significant duplication at each stage. This causes data fragmentation as multiple copies and versions are scattered across different processing steps, workgroups, and storage locations, further limiting cross-domain applications. In this paper, we propose a unified point cloud data management (PCDM) approach that supports the principles of findability, accessibility, interoperability, and reusability (FAIR) across domains at scale. The proposed approach aims to support diverse point cloud retrieval for cross-domain analysis by leveraging a single, reusable PCDM system built on a shared data model. Our approach improves on existing frameworks and provides a foundation for point cloud data management and data spaces.

1. Introduction

Digital modeling of environments using high-resolution geospatial data, including 3D point clouds, has experienced substantial growth over the past decade, driven by advances in sensor technology, computer vision, and image processing. These innovations have accelerated the adoption of realistic 3D models across a wide range of scientific domains. The virtual representation of complex scenes and objects using 3D point clouds has become widely accepted across fields such as autonomous driving, robotics (Básaca-Preciado et al., 2014), structural health monitoring (Merkle et al., 2023), and medical scanning (Rodriguez-Quinonez et al., 2013).

A point cloud consists of a set of data points that represent the most fundamental geometric entity within a three-dimensional coordinate system. Each point is defined by its unique (X, Y, Z) location and does not possess scale, width, length, or height; rather, they collectively describe the surface of an object (Wang and Kim, 2019). Additionally, these spatially independent data nodes can store supplementary information, including normal vectors, curvature, neighborhood relationships, color, texture, material properties, and other relevant attributes. Due to their ability to capture a wide spectrum of surfaces at various scales, point clouds are an effective data modality for applications in science and engineering, representing both complex and simple object surfaces.

Cross-domain analytics on point cloud is constrained by how data are acquired, stored, and semantically annotated. Heterogeneous acquisition pipelines using different sensors and techniques produce data with varying densities, noise characteristics, reference frames, and formats, which further complicates the use of point clouds across multiple domains (Remondino et al., 2018). Additionally, semantic labeling is often performed based on domain-specific analyses, using divergent class vocabularies, which fragments the semantic layer and hinders interoperability and reuse (Shao et al., 2019).

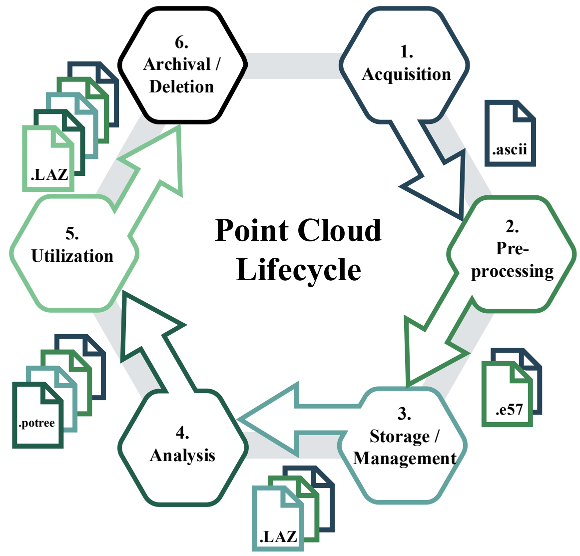

Figure 1. Point cloud lifecycle illustrating data volume growth and how multiple file formats emerge across the various stages.

While there are some semi-standardized representations for point clouds, such as LAS, the variety of sensors and semantics challenges the fusion of multisensor and cross-domain data in a unified data space, such as a digital twin platform. This work views this problem from a data lifecycle perspective, aiming to provide a holistic perspective and insights into point cloud data management and data spaces (Wilkinson et al., 2016). The primary focus is to conceptualize a unified data management approach that supports the principles of findability, accessibility, interoperability, and reusability (FAIR) across domains at scale.

2. Point Cloud Lifecycle

The data lifecycle for point clouds includes generation, processing, management, analysis, utilization, and archival or deletion across various domains (Figure 1). While individual stages may be adapted to specific use cases, the lifecycle generally follows a familiar sequence consistent with other data forms.

Data acquisition of point clouds typically involves measuring surfaces using remote sensor techniques. These include active methods, such as 3D laser scanners, and passive approaches, using photogrammetry and structure-from-motion, or a combination of both, to generate point clouds. The quality of the raw data, specifically its precision, accuracy, coverage, and resolution, affects subsequent processes. It is therefore essential to ensure that the data meets the requirements of the intended applications while maintaining quality and avoiding redundancy.

Pre-processing of the raw data involves a series of steps, including filtering noise and outliers, alignment of different point clouds into a unified coordinate system using registration, and, in specific applications, subsampling to reduce the number of points to make the processing more efficient, among others.

Data management of point clouds is modeled in fixed-size record-oriented data structures, sliced into chunks, and stored in a variety of file formats (e.g., $.LAS/LAZ$, $.e57$, $.potree$, $.ascii$, $.txt$, or $.pcd$), local and on servers. The data is processed using tools such as PDAL and LAStools, and visualized with tools such as CloudCompare and Potree web viewer.

Analysis of point clouds is a domain-specific process, which depends on the intended application. This may include semantic enrichment through segmentation and classification for building information modeling, as well as the extraction of geometric models such as meshes, surface representations, and finite element models, or structural health assessments, including deformation and crack detection.

Utilization: the domain-specific information derived from point cloud analysis is communicated through visualization tools, monitoring platforms, and simulation environments to support decision-making. These outputs also facilitate advanced interactive analyses such as change detection, asset management, and inventory analysis.

Archiving point cloud data involves migrating files from primary storage to long-term archival storage, where access controls are implemented to regulate retrieval. To address compliance and GDPR requirements, value-based data deletion is recommended at the end of the point cloud data lifecycle, ensuring that only data necessary for retention purposes is preserved and all redundant or obsolete data is securely removed.

The state-of-the-art lifecycle of point clouds, as shown in Figure 1, demonstrates growth in data along the cycle due to the lack of standardization of point clouds. Data duplication across layers occurs when point cloud files are converted to the dedicated file format required for each stage. For example, to perform a neighborhood analysis and visualize a $.laz$ point cloud using the Python library Open3D, the data must be converted to the $.pcd$ format to fit the tooling. On the other hand, if the same data has to be published on a web platform that supports Potree web-viewer, the data has to be converted into the $.potree$ file format. As shown in the diagram, the data at the end of the cycle grows to approximately five times the volume of the captured point cloud. Hence, there is a need to investigate an efficient approach to unify, standardize, and govern point cloud data that supports diverse workloads.

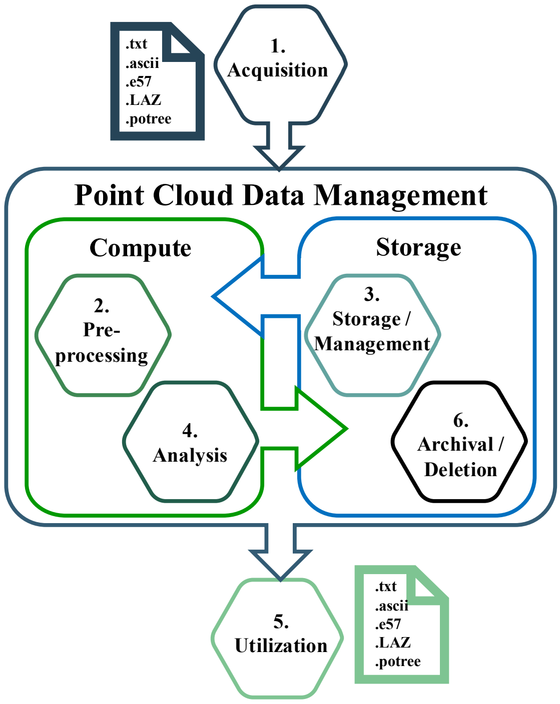

Figure 2. Point cloud lifecycle processing integrated in the proposed point cloud data management.

3. Point Cloud Data Management

There are numerous point cloud data management solutions that rely on various database management system architectures. However, they often exhibit limitations such as subpar scalability, data duplication and amplification, limited indexing capabilities, constrained accessibility due to opaque storage, or inefficient interoperability caused by transformation overhead. Hence, we propose a unified, decoupled storage and compute framework for point cloud data management that supports scalable processing and analysis, as shown in Figure 2.

In preliminary work, we explored a novel architecture coined *Lakehouse*, that extends and specifies a NoDB approach tailored for modern cloud computing environments (Teuscher and Werner, 2025). It offers scalability and flexibility through the disaggregation of storage and compute, which can run across various environments, including single-node, distributed, and cluster environments, as well as on-premises and cloud environments. Additionally, it relies on Apache Arrow as the self-describing in-memory format, which is gaining increasing support in the broader analytical ecosystem, allowing data to be exchanged with zero-copy and translation overhead even across applications and system boundaries.

Building upon the insights and lessons learned from this work, we envision integrating and supporting multiple point cloud representations, such as raw points, voxels, meshes, and graphs, within a unified data structure and management system. The primary objective of this endeavor is to address the diverse retrieval needs of multiple users from a single data model through a reusable data layer. Furthermore, a key premise is to integrate results from analytical workflows into the system without duplicating data. This necessitates providing an extensible and modifiable data model to overcome the limitations of existing formats.

The proposed PCDM *Lakehouse* architecture offers several advantages and functionalities over traditional point cloud data management stacks. The storage component stores point clouds as objects in scalable, S3-compatible storage or cloud blobs, decoupled from compute. This allows storage to scale independently to handle growing and large datasets without incurring unnecessary increases in compute resources, which can be very expensive. We propose using tabular formats such as Lance (Pace et al., 2025) or Apache Iceberg[1] as a columnar storage layer that stores both point clouds and metadata together, optimized for big-data processing. They support schema evolution and ACID-like transactions for inserting, deleting, and updating data, with conflict resolution via multi-version concurrency control.

The metadata layer is central to enabling cross-domain operations and clear ontology definition by providing shared semantics, structure, and context across point cloud datasets, effectively acting as a “semantic glue” for domain-specific use (Da Silva et al., 2014). However, there is a downside to endlessly growing metadata: in cross-domain settings, extensive definitions and overlapping classes can lead to ontology bloating, where many nearly identical classes and relations make the system hard to understand, evolve, and maintain. To mitigate this effect, the PCDM will enforce metadata quality and lifecycle rules to govern operations and prevent unbounded metadata growth.

Another premise of the proposed architecture is that sharing point cloud processing pipelines is recommended as scripts, while the data source is envisioned to be shared through public or authenticated access of the PCDMs decoupled storage layer. This ensures that different domains compute over the same *Lakehouse* storage with aligned semantics and workflows, thereby enabling cross-domain fusion, comparison, and joint analysis at lower cost, with less duplicated compute, and with more semantically sound results. To facilitate this, workflows may require some adaptation to shift the data preprocessing and retrieval to a declarative approach. In case this is not feasible or suitable, creating synopses or services on top is possible.

4. Cross-Domain Applications

In this section, we outline several cross-domain applications exemplifying the implications and potential benefits of integrating them into the proposed system.

4.1 Urban Scene Update

One application that can significantly benefit from the proposed PCDM architecture is urban scene update (Stilla and Xu, 2023). As different domains focus on measuring distinct objects of interest at varying resolutions, the PCDM can perform voxel-based weighted point cloud fusion (Poku-Agyemang and Reiterer, 2025), enabling fine-grained updates. Scene analysis can then be implemented as an event-driven mechanism that automatically refreshes the model whenever a new epoch is uploaded to the PCDM. The voxel-based fusion analysis assesses the quality of the data source and assigns a Boolean value indicating whether to retain or remove each voxel, based on its relevance and importance at the epoch. In the classical approaches presented in the paper, the analysis results in a new copy of the point clouds for each analysis, thereby significantly increasing the dataset with each new epoch.

[1] `https://iceberg.apache.org/`

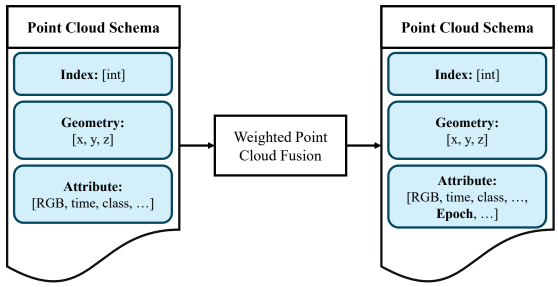

Figure 3. Point cloud schema evolution by adding a new attribute without the need to copy the original data.

The voxel-based scene update can be performed directly on the point clouds stored in the PCDM, where a voxel query retrieves the relevant data into the analysis pipeline, and the resulting labels are appended as attributes for each epoch. The schema is extended with temporal information indicating, for each upload, which points are flagged as inliers or outliers, allowing multiple point clouds of the same region to be incorporated without data duplication, as shown in Figure 3. The scripts for the analytical pipeline can be shared and stored as metadata, enabling domain-specific reusability and parallel, scalable compute.

4.2 Semantic Point cloud analysis

Semantically classified point clouds are used across many domains for class–object analysis to reveal patterns and events of interest (Xie et al., 2020). For instance, in (Letard et al., 2024), the authors propose an explainable machine learning approach for multispectral point cloud classification based on bispectral lidar data covering large coastal and fluvial environments. The resulting topo-bathymetric lidar datasets for coastal environments comprise 13 detailed classes, including vegetation, ground, artificial ground, buildings, sandy seabed, and pebble, among others, which can be used for further domain-specific analysis. These analyses often require class-specific point cloud queries that are not well supported by traditional point cloud data management infrastructures.

Classical workflows typically load the entire point cloud into memory and then filter to segment the desired classes within the spatial bounds of interest, which can be computationally expensive and become a bottleneck for large-scale aerial analyses. The proposed PCDM enables fast and efficient spatial data retrieval by supporting SQL-like queries for direct class-specific point selection from the database. From the PCDM, segments of a large dataset can be queried using either object-specific or spatial-specific queries and fed directly into the analytical pipeline for the analysis from the same point cloud dataset as shown in Figure 4. Additionally, these processes can be done simultaneously on different compute units attached to the storage unit of the *Lakehouse* architecture.

4.3 Point Cloud Visualization

Visualization has long been a main bottleneck for large-scale point cloud rendering, primarily due to data throughput and

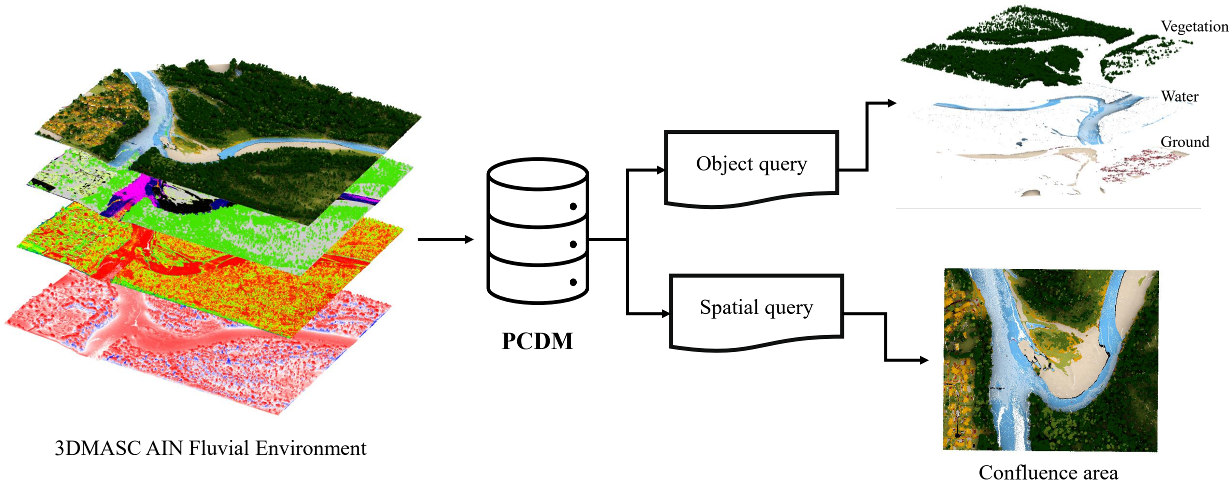

Figure 4. Data retrieval for domain specific analysis examplified with AIN fluvial environment dataset with of semantic information (11 classes, PCV luminance and RGB-Color information) of a portion of the Ain River in South-Eastern France near its confluence with the Rhone River .

memory-bandwidth pressure along the full pipeline—from storage to GPU VRAM and into the rasterization stage. To interactively explore multimodal and multitemporal point-cloud data covering large-scale areas, such as the TUM2TWIN dataset (Wysocki et al., 2026), poses significant challenges. The dataset comprises terrestrial laser scanning (TLS), mobile laser scanning (MLS), UAV-based laser scanning (ULS), and aerial laser scanning (ALS) data over the same urban area, all in a common reference frame, making it computationally demanding to visualize on consumer-grade computers.

Traditional viewers such as Potree address this by relying on LOD structures that are typically precomputed as a separate, offline preprocessing step, which must be completed before interactive visualization is possible (Schütz et al., 2024). This process must be repeated whenever the data is expanded or modified, rendering it inefficient for dynamic or frequently updated use cases. In the proposed PCDM architecture, a combination of continuous and discrete levels of detail (LOD) is precomputed once and stored as additional attributes, effectively decoupling LOD generation from the rendering client (Figure 5). The pre-computed LODs within the PCDM facilitate large-scale rendering by turning it into a query-driven, tiled, or spatial-chunk retrieval problem, replacing client-side, bandwidth- and CPU-bound LOD generation with scalable queries that keep the GPU pipeline fed with the appropriate LOD.

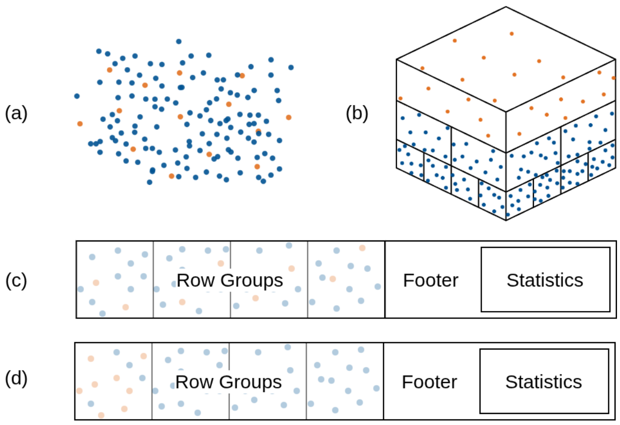

Figure 5. Point cloud data organization based on importance (orange) for visualization: a) point set with important points, b) hierarchical space partitioning incorporating importance, c) importance locality in files without partitioning, d) importance locality with partitioning.

While we demonstrated the feasibility of this approach on static datasets through appropriate partitioning, integrating it into tabular formats and dynamic datasets is ongoing research. For example, tabular formats offer some basic capabilities for partitioning data. However, spatial partitioning based on importance semantics needs further exploration and work to become a first-class citizen. A promising approach, conceptualized in prior work (Teuscher and Werner, 2024), is to adapt a Log-Structured Merge-Tree (LSMT) that understands spatial and importance semantics, thereby iteratively refining partitioning throughout the whole lifecycle.

5. Discussion

The proposed PCDM improves upon existing point cloud management systems by bringing big data and data fusion concepts into point cloud processing and analysis. The key advantage over traditional and classical point cloud management systems is the ability to modify the schema and append analytical results without duplicating data. A major challenge in using point cloud data across several scientific domains is the repeated duplication of data for each analytical process, since different spatial resolutions, segments, and attributes of the point cloud dataset are required for different analyses. This can lead to significant data growth and fragmentation across the various stages of the analysis, from data acquisition to information extraction.

As demonstrated in the three use cases, modifying the point cloud schema to store results for domain-specific analysis helps prevent multiple copies of the data from being created for each application. Data retrieval using SQL-like queries for object-specific or spatially bounded point cloud segments directly improves the efficiency of domain-specific analytics. Precomputed partitions of the point cloud, based on defined levels of detail, enable efficient rendering, thereby enhancing visualization and interactive exploration. The PCDM also supports simultaneous visualization and analytics without requiring the data to be transformed into compatible formats, enabling tightly coupled workflows over the same underlying storage.

The proposed infrastructure provides a unified logical point cloud management layer over distributed and localized data storage, enabling cross-domain applications and serving as a foundation for scientific data spaces for collaborative research. Without forcing a single monolithic repository or requiring physical data exchange among point cloud research communities, the PCDM can be used to share vocabularies, metadata models, and mappings that bridge domain-specific schemas and support interoperability frameworks grounded in shared rules and governance. By decoupling data storage and compute in the proposed infrastructure, point clouds can be reused across multiple data spaces, reducing redundant integration effort while opening up new use cases beyond the point cloud research communities.

The proposed PCDM demonstrates strong potential to close the gap in cross-domain collaboration; however, several limitations could hinder this process. To ensure effective cross-domain use of the PCDM, the ontology and domain lexicons must be clearly defined and rigorously governed. Designing a single cross-domain ontology is inherently difficult, as different domains often use similar terms with different meanings or different terms for the same concept. This semantic heterogeneity can lead to high alignment costs, challenges in integrating the ontology with existing metadata systems, and additional usability overhead for non-expert users.

The next steps envisioned are to build out a prototype system that incorporates real-time ingestion, semantic augmentation, transactions, and workload-aware partitioning. First steps toward the latter have been proposed in recent work for incremental reclustering in cloud data warehouses (Liu et al., 2026). Of particular interest in this regard is a thin, decoupled management layer based on a LSMT, with a learned approach reinforced by actual workloads driving continuous, adaptive repartitioning in the background.

6. Conclusion

In this paper, we propose a point cloud data management framework based on *Lakehouse* a unified, decoupled storage and compute architecture. This approach enables scalable processing and analysis of point cloud data, enhancing efficiency and flexibility, as demonstrated in selected cross-domain applications. The proposed framework is designed to support diverse workloads, paving the way for interdisciplinary research, parallel analysis, and both real-time and batch processing of large point cloud datasets. The presented architecture demonstrates advantages over traditional and classical point cloud management systems. Developed with data protection and security in mind, the framework can be deployed across a range of environments, including single-node, distributed, clustered, and cloud settings. Ultimately, this supports maximizing the benefits of spatial understanding of the environment by using point clouds across various disciplines and applications.

References

Básaca-Preciado, L. C., Sergiyenko, O. Y., Rodríguez-Quinonez, J. C., García, X., Tyrsa, V. V., Rivas-Lopez, M., Hernandez-Balbuena, D., Mercorelli, P., Podrygalo, M., Gurko, A. et al., 2014. Optical 3D laser measurement system for navigation of autonomous mobile robot. *Optics and Lasers in Engineering*, 54, 159–169.

Da Silva, J. R., Ribeiro, C., Lopes, J. C., 2014. Ontology-based multi-domain metadata for research data management using triple stores. *Proceedings of the 18th International Database Engineering & Applications Symposium*, 105–114.

Letard, M., Lague, D., Le Guennec, A., Lefèvre, S., Feldmann, B., Leroy, P., Girardeau-Montaut, D., Corpetti, T., 2024. 3DMASC: Accessible, explainable 3D point clouds classification. Application to bi-spectral topo-bathymetric lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 207, 175-197.

Liu, Y., Zhou, R., Yan, J., Zhang, H., 2026. Workload-Aware Incremental Reclustering in Cloud Data Warehouses. *arXiv preprint arXiv:2602.23289*.

Merkle, D., Solass, J., Schmitt, A., Rosin, J., Reiterer, A., Stolz, A., 2023. Semi-automatic 3D crack map generation and width evaluation for structural monitoring of reinforced concrete structures. *Journal of Information Technology in Construction*, 28.

Pace, W., She, C., Xu, L., Jones, W., Lockett, A., Wang, J., Shah, R., 2025. Lance: Efficient random access in columnar storage through adaptive structural encodings.

Poku-Agyemang, K. N., Reiterer, A., 2025. Weighted multiple point cloud fusion. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 93(1), 65–78.

Remondino, F. et al., 2018. Point cloud acquisition and structuring. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5, 251–258.

Rodriguez-Quinonez, J. C., Sergiyenko, O., Gonzalez-Navarro, F. F., Basaca-Preciado, L., Tyrsa, V., 2013. Surface recognition improvement in 3D medical laser scanner using Levenberg–Marquardt method. *Signal Processing*, 93(2), 378–386.

Schütz, M., Herzberger, L., Wimmer, M., 2024. SimLOD: Simultaneous LOD generation and rendering for point clouds. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 7(1), 1–20.

Shao, T. et al., 2019. Semantic labeling and instance segmentation of 3D point clouds. *IEEE Transactions on Visualization and Computer Graphics*, 25(5), 1839–1853.

Stilla, U., Xu, Y., 2023. Change detection of urban objects using 3D point clouds: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197, 228–255.

Teuscher, B., Werner, M., 2024. Pc-lmt: The point cloud log merge tree for the helena point cloud database. *Proceedings of the 12th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, BigSpatial ’24, Association for Computing Machinery, New York, NY, USA, 1–9.

Teuscher, B., Werner, M., 2025. Point Cloud Data Management for Analytics in a Lakehouse. *AGILE: GIScience Series*, 6, 1–7. https://agile-giss.copernicus.org/articles/6/47/2025/.

Wang, Q., Kim, M.-K., 2019. Applications of 3D point cloud data in the construction industry: A fifteen-year review from 2004 to 2018. *Advanced engineering informatics*, 39, 306–319.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1–9.

Wysocki, O., Schwab, B., Biswanath, M. K., Greza, M., Zhang, Q., Zhu, J., Froech, T., Heeramaglore, M., Hijazi, I., Kanna, K. et al., 2026. TUM2TWIN: Introducing the large-scale multimodal urban digital twin benchmark dataset. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232, 810–830.

Xie, Y., Tian, J., Zhu, X. X., 2020. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geoscience and remote sensing magazine*, 8(4), 38–59.