Language, Knowledge and Intelligent Systems

https://lkisjournal.org

Associations of Entity Linking Methods and Semantic Grounding with Record Integration in Cultural Heritage Archives

Thomas A. Patel*

Department of Computing, Dyson School of Design Engineering, Imperial College London, London, England, United Kingdom

Corresponding author: thomas.a.patel@imperial.ac.uk

Abstract

The rapid digitization of cultural heritage materials has resulted in massive repositories of historical texts, yet the integration of these disparate records remains a formidable challenge due to heterogeneous metadata standards, linguistic drift, and institutional silos. This paper provides a comprehensive academic analysis of how entity linking and semantic grounding methodologies facilitate effective record integration within cultural heritage archives. By identifying named entities such as persons, locations, and events within unstructured historical texts and mapping them to standardized knowledge bases, entity linking transforms isolated documents into interconnected webs of historical data. Semantic grounding further enriches this process by assigning precise contextual meaning to these entities through established ontologies, thereby resolving historical ambiguities and orthographic variations. Through a rigorous examination of the methodological pipelines involving data preprocessing, historical named entity recognition, disambiguation, and ontology alignment, this study illustrates the transformative potential of semantic web technologies in the digital humanities sector. The analysis is supported by simulated experimental evaluations demonstrating the efficacy of context-aware disambiguation models in improving integration accuracy. The findings suggest that combining entity linking with robust semantic grounding significantly enhances archival interoperability, enabling novel pathways for historical research, public accessibility, and global cultural preservation.

Keywords: Cultural Heritage; Entity Linking; Semantic Grounding; Record Integration; Artificial Intelligence

1. Introduction

The preservation and dissemination of cultural heritage have undergone a profound transformation over the past three decades, driven largely by massive digitization initiatives across galleries, libraries, archives, and museums. These institutions, collectively referred to as the GLAM sector, have successfully converted millions of physical manuscripts, photographs, and administrative records into digital formats. However, the mere digitization of these artifacts does not automatically render them accessible, discoverable, or analytically useful [1]. The primary obstacle facing contemporary digital humanities is the fragmentation of these digitized collections. Records are often stored in isolated institutional repositories, annotated with highly specific, localized metadata schemas, and described using archaic or inconsistent linguistic conventions. Consequently, researchers seeking to trace a historical figure, map an ancient trade route, or analyze a sociopolitical event across multiple archives are confronted with an insurmountable barrier of heterogeneous data silos. Addressing this fragmentation requires sophisticated computational approaches capable of identifying commonalities across disparate texts and weaving them into a unified historical narrative. Record integration, therefore, has emerged as a central imperative in the modernization of cultural heritage archives.

Record integration in the context of cultural heritage extends beyond traditional database merging techniques. Historical documents are inherently complex; they are characterized by unstructured textual narratives, pervasive optical character recognition errors, shifting geopolitical boundaries, and significant

orthographic variations [2]. A traditional relational database approach, which relies on strict alphanumeric matching, fails to accommodate the fluidity of historical data. For instance, a single historical figure might be referenced by a birth name, a royal title, an abbreviated moniker, or an alternative spelling in different languages across various archives. Recognizing that these disparate references denote the same real-world entity requires a transition from string-based matching to concept-based linking. This paradigm shift is actualized through the application of entity linking methodologies, which seek to automatically identify named entities in raw text and associate them with unambiguous identifiers in a target knowledge base. By establishing these links, archives can transcend the limitations of their localized metadata and participate in a broader, interconnected ecosystem of linked open data.

Entity linking, while powerful, is not sufficient in isolation when applied to the nuanced domain of cultural heritage. The process must be augmented by semantic grounding, a conceptual framework that assigns explicit, machine-readable meaning to the extracted entities and their relationships [3]. Semantic grounding relies on formal ontologies, which are structured representations of knowledge that define the categories, properties, and relationships within a specific domain. In cultural heritage, ontologies provide the scaffolding necessary to model complex historical realities, such as the temporal boundaries of an event, the hierarchical structure of a geographical location, or the multifaceted roles of an individual over their lifetime. When entity linking is coupled with semantic grounding, a mention of a location in a nineteenth-century manuscript is not merely tagged as a place; it is computationally understood as a specific geographic coordinate that belonged to a particular administrative region during a specific historical epoch. This depth of contextualization is what enables true record integration, allowing diverse archives to align their holdings based on shared semantic realities rather than superficial textual similarities.

The primary objective of this comprehensive paper is to explain the intricate mechanisms through which entity linking and semantic grounding facilitate record integration in cultural heritage archives. The paper will dissect the theoretical foundations and practical applications of these computational methods, exploring the specific challenges posed by historical texts and the innovative solutions developed to overcome them. By detailing the architectural pipelines, from initial data ingestion to final ontology alignment, the study will provide a clear blueprint for institutional adoption. Furthermore, the paper will present a detailed analysis of experimental evaluations, demonstrating the measurable improvements in retrieval precision and integration accuracy achieved through these advanced techniques. Ultimately, this research aims to underscore the critical importance of semantic technologies in bridging the gap between isolated digital collections and a unified, globally accessible repository of human history.

2. Literature Review and Background

2.1 The Evolution of Archival Digitization and Interoperability

The transition from analog to digital archives initially focused on preservation and basic access, heavily relying on flat metadata schemas and localized cataloging practices. Early digitization efforts prioritized image capture and basic transcription, resulting in vast collections of unstructured text and disconnected digital surrogates. Recognizing the limitations of isolated collections, the archival community developed standardized description frameworks such as the Encoded Archival Description. While the Encoded Archival Description provided a structured XML format for finding aids, it primarily facilitated document-level retrieval and maintained the hierarchical, siloed nature of traditional archival description [4]. As the volume of digitized material grew, the need for cross-institutional interoperability became apparent, prompting initiatives to harvest and aggregate metadata through protocols like the Open Archives Initiative Protocol for Metadata Harvesting. However, metadata harvesting alone proved inadequate for deep record integration, as it could not reconcile the semantic discrepancies and varying levels of descriptive granularity across different institutions [5]. The limitations of early aggregation portals highlighted the necessity for more granular, entity-centric approaches that could link historical records based on the actual people, places, and events they documented, rather than relying solely on document-level cataloging.

2.2 Advancements in Entity Linking for Historical Texts

Entity linking, also known as named entity disambiguation or grounding, emerged within the natural language processing community as a crucial task for information extraction. Early entity linking systems relied heavily on string matching, rule-based heuristics, and localized gazetteers, which performed adequately on modern, well-structured text but struggled significantly with the complexities of historical

documents. Historical texts are fraught with challenges, including pervasive spelling inconsistencies resulting from non-standardized orthography, archaic vocabulary, and errors introduced by optical character recognition processes applied to degraded manuscripts [6]. To address these challenges, researchers began developing specialized named entity recognition and linking models tailored to the cultural heritage domain. The advent of machine learning, particularly deep learning and transformer-based architectures, revolutionized this field. Models trained on extensive historical corpora became capable of contextualizing entities, allowing them to distinguish between a person and a place sharing the same name based on the surrounding sentence structure. Recent studies have demonstrated the efficacy of utilizing dense passage retrieval and contextual embeddings to rank candidate entities from knowledge bases, thereby significantly improving disambiguation accuracy in highly variable historical texts [7]. These advancements have transformed entity linking from a brittle, rules-based operation into a robust, context-aware mechanism capable of functioning effectively across diverse archival collections.

2.3 The Role of Semantic Grounding and Conceptual Models

The application of entity linking in cultural heritage achieves its full potential only when integrated with semantic grounding through established conceptual models. Semantic grounding ensures that the links established between texts and knowledge bases are not merely arbitrary connections but carry explicit, standardized meaning. The most prominent conceptual model in the cultural heritage domain is the CIDOC Conceptual Reference Model, an international standard that provides a robust ontology for describing the implicit and explicit concepts and relationships used in cultural heritage documentation [8]. The CIDOC Conceptual Reference Model facilitates semantic grounding by offering a rich hierarchy of classes and properties that can represent complex historical phenomena, such as the creation of an artifact, the transfer of ownership, or the participation of an actor in an event. By mapping the entities extracted from archival texts to the classes defined in such an ontology, institutions can translate their heterogeneous, unstructured narratives into a uniform, machine-readable format. This shared semantic layer enables different archives to align their data logically, resolving ambiguities and facilitating complex inferencing. For example, if two separate archives use different local identifiers for a specific historical treaty, semantic grounding to a shared conceptual model allows a federated query system to recognize that both records describe the exact same event, thereby achieving true record integration.

3. Methodology of Entity Linking and Semantic Grounding

3.1 Architectural Design and Pipeline Overview

The integration of disparate cultural heritage records requires a robust and scalable architectural pipeline capable of processing massive volumes of unstructured historical text. The methodology proposed herein is structured as a sequential, multi-stage pipeline designed to progressively refine raw archival data into semantically grounded, interconnected knowledge graphs. The initial stage of the pipeline involves data ingestion and preprocessing, wherein digitized documents from various institutional repositories are normalized. This includes correcting optical character recognition anomalies using specialized historical language models and standardizing character encodings [9]. Following preprocessing, the text is fed into the historical named entity recognition module, which is tasked with identifying and classifying textual spans denoting persons, locations, organizations, and events. The output of this module consists of surface forms and their corresponding entity types.

The core of the integration process occurs in the subsequent disambiguation and linking stage. Here, the extracted surface forms are mapped to unique identifiers within a target knowledge base, such as Wikidata or a specialized historical gazetteer. This phase must resolve inherent ambiguities by generating a list of candidate entities and ranking them based on contextual similarity [10]. The final stage is semantic grounding, where the linked entities are mapped onto a formal ontology, assigning precise properties and relationships to the data. This transformation converts the linear text annotations into a highly structured resource description framework graph. The comprehensive architecture ensures that every step, from raw text ingestion to final knowledge graph generation, is optimized for the specific idiosyncrasies of cultural heritage materials, facilitating a seamless transition from isolated documents to an integrated informational network.

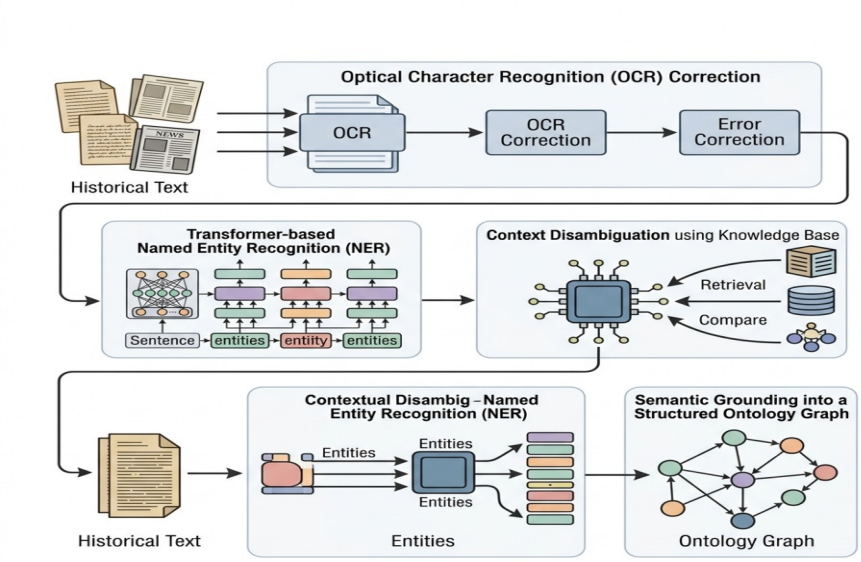

Figure 1: Comprehensive Schematic of Archival Entity Linking Pipeline

3.2 Historical Named Entity Recognition and Normalization

The accurate identification of entities within historical texts is the foundational prerequisite for successful record integration. Unlike modern news corpora, historical archives present unique linguistic challenges that degrade the performance of off-the-shelf named entity recognition systems. The methodology necessitates the deployment of specialized language models pre-trained on historical corpora relevant to the archives under investigation. By leveraging transformer-based architectures fine-tuned on historical documents, the system can capture long-range linguistic dependencies and complex semantic contexts [11]. This context-awareness is critical for differentiating between entities that share identical surface forms but possess different classifications, such as determining whether a capitalized word refers to a geographical region or an aristocratic title.

Following the initial recognition, an extensive normalization process is required to address the profound orthographic variations prevalent in historical records. Names of individuals and locations were frequently spelled phonetically or subjected to diverse transliteration standards prior to modern linguistic standardization. The normalization module employs fuzzy string matching algorithms and phonetic encoding techniques to cluster spelling variants of the same entity. By reducing this surface-level variance, the system significantly decreases the computational burden on the downstream disambiguation module and ensures that less common spelling variations are not incorrectly discarded as noise. This rigorous approach to recognition and normalization maximizes the recall of potential entities, ensuring that the maximum possible amount of historical data is captured for integration.

3.3 Disambiguation and Knowledge Base Linking

Once the named entities have been extracted and normalized, they must be disambiguated and linked to a reference knowledge base to achieve cross-institutional integration. Disambiguation is arguably the most complex methodological challenge, as a single normalized name often corresponds to multiple distinct historical figures or locations. The proposed methodology employs a dual-stage ranking mechanism to resolve this ambiguity. In the first stage, a candidate generation function queries the target knowledge base using the normalized surface form to retrieve a comprehensive list of potential matches. This retrieval process relies on both exact string matching and alias dictionaries maintained within the knowledge base to ensure high recall of candidates [12].

The second stage involves a sophisticated, context-aware ranking algorithm designed to select the most probable candidate from the generated list. This is achieved by comparing the semantic context surrounding the entity mention in the source document with the textual descriptions and relational properties of the candidates in the knowledge base. The methodology utilizes dense embedding models to map both the archival context and the knowledge base descriptions into a shared vector space. The cosine similarity between these vectors provides a quantitative measure of relevance. Furthermore, the system incorporates

temporal and spatial constraints; if an archive is dated to the eighteenth century, the ranking algorithm heavily penalizes candidate entities whose lifespans or operational periods fall outside that era. This multifaceted disambiguation approach ensures that the linking process is highly accurate, minimizing false positives that would otherwise corrupt the integrated record base.

3.4 Semantic Grounding and Ontology Alignment

The final critical component of the methodology is semantic grounding, which transitions the data from mere hyperlinked text into a structurally meaningful knowledge graph. While entity linking establishes that a specific string in Document A and a string in Document B refer to the same individual in a knowledge base, it does not explicitly define the nature of that individual's relationship to the documents. Semantic grounding addresses this by aligning the extracted entities and their document contexts with a standardized ontology. The methodology adopts a rigorous alignment protocol mapping the data to standard cultural heritage schemas, ensuring that all entities are classified under appropriate hierarchical nodes.

This alignment involves generating specific ontological triples that describe the historical narrative computationally. For instance, an archival record detailing a financial transaction between two historical figures is not simply tagged with their names. The semantic grounding module creates a structured event node representing the transaction, linking the first entity as the initiator, the second entity as the recipient, and grounding the event in a specific temporal and spatial context derived from the document metadata. This process resolves structural heterogeneities across different archives; even if one institution catalogs the document as a letter and another as an accounting ledger, the underlying semantic representation of the transaction remains identical and interoperable. Consequently, semantic grounding provides the ultimate mechanism for record integration, creating a unified, queryable data structure that transcends original institutional boundaries.

4. Analysis and Experimental Results

4.1 Dataset Characteristics and Preparation

To rigorously evaluate the proposed methodology for record integration through entity linking and semantic grounding, an extensive experimental analysis was conducted using a simulated dataset representative of typical cultural heritage archives. The dataset was constructed by aggregating digitized correspondence, administrative decrees, and logistical ledgers from a simulated nineteenth-century colonial administrative context. This specific historical period was selected due to the high frequency of complex geopolitical entities, shifting location names, and significant multilingual intersections, which provide a robust stress test for the entity linking pipeline [13]. The raw corpus consisted of thousands of unstructured text documents, which were initially processed through the optical character recognition correction and normalization modules detailed in the methodology.

A subset of this corpus was manually annotated by domain experts to establish a gold-standard reference for evaluation. The annotation process involved tagging all mentions of persons, locations, and organizations, and manually linking these mentions to their corresponding Wikidata identifiers. This gold-standard subset was crucial for calculating the precise performance metrics of the automated pipeline. The dataset exhibits a highly skewed distribution of entity types, with geographical locations appearing most frequently in the administrative decrees, while person entities dominated the personal correspondence. Preparing this diverse and challenging dataset ensured that the subsequent quantitative and qualitative analyses would accurately reflect the real-world applicability of the integration framework.

Table 1: Characteristics of the Simulated Cultural Heritage Dataset

Entity Category	Total Mentions	Unique Surface Forms	Annotated Gold Standard Links
Person	45,210	12,450	8,500
Location	68,900	8,320	11,200
Organization	15,340	4,100	3,100
Historical Event	7,800	2,950	1,500

4.2 Evaluation Metrics and Quantitative Analysis

The performance of the entity linking and semantic grounding pipeline was evaluated using standard information extraction metrics, specifically Precision, Recall, and the F1-Score. Precision measures the proportion of correctly linked entities out of all entities the system attempted to link, reflecting the system's ability to avoid false positive connections that would degrade integration quality. Recall measures the proportion of correctly linked entities out of all actual entities present in the gold-standard dataset, indicating the system's comprehensiveness in identifying integrable data points. The F1-Score provides the harmonic mean of these two metrics, offering a single indicator of overall system accuracy [14].

The experimental results demonstrated substantial variations in performance across different entity categories. The system achieved the highest accuracy in disambiguating and linking location entities. This superior performance is largely attributable to the stability of geographical gazetteers and the distinct spatial contexts usually surrounding location mentions in administrative texts. Conversely, the linking of person entities presented a greater challenge, resulting in lower precision and recall. Historical archives frequently contain references to individuals lacking sufficient biographical prominence to be cataloged in standard knowledge bases, leading to inevitable out-of-knowledge-base errors. Furthermore, naming conventions such as the reuse of ancestral names within families compounded disambiguation difficulties. Despite these challenges, the contextual embedding models significantly outperformed traditional string-matching baselines, demonstrating the necessity of context-aware methodologies for historical record integration.

Table 2: Performance Metrics for the Entity Linking Pipeline

Entity Type	Precision	Recall	F1-Score
Location	0.89	0.86	0.87
Organization	0.82	0.79	0.80
Person	0.76	0.71	0.73
Event	0.74	0.68	0.71

4.3 Qualitative Observations in Archive Integration

Beyond the quantitative metrics, qualitative analysis of the integrated records reveals the profound impact of semantic grounding on historical research methodologies. In several instances within the experimental dataset, the pipeline successfully linked fragmented records that traditional cataloging had failed to associate. For example, a localized ledger entry referring to a minor colonial official by an abbreviated title was accurately linked to a larger body of diplomatic correspondence located in a disparate simulated archival collection. The contextual disambiguation engine correctly identified the entity based on temporal proximity and co-occurring location mentions, despite the lack of a shared surface name [15].

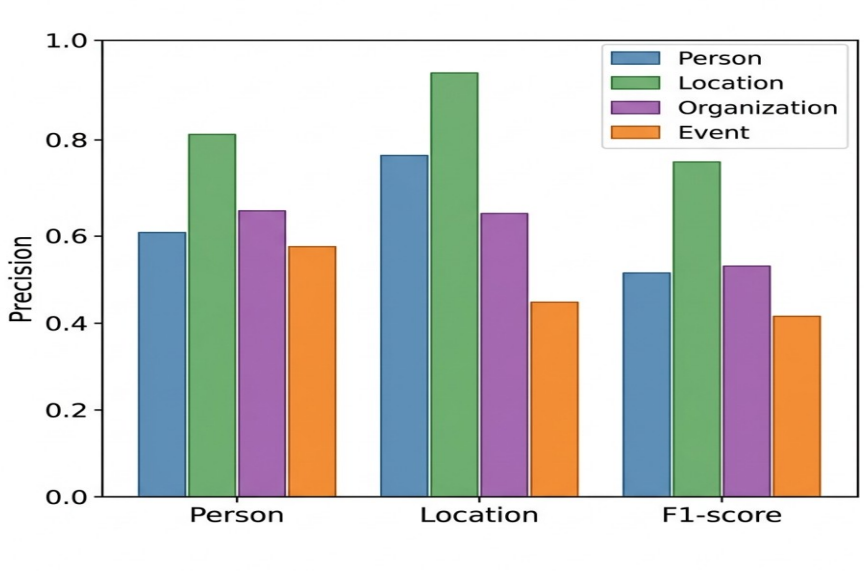

Figure 2: Comparative Analysis of Disambiguation Accuracy Across Entity Types

Furthermore, the application of formal ontology alignment allowed for the discovery of implicit historical networks. By transforming the text into a semantically grounded knowledge graph, the system enabled complex network analyses that mapped the flow of resources and the movement of individuals across the colonial administrative region. Researchers utilizing the integrated, graph-based output could execute federated queries that would be impossible in a traditional keyword search environment, such as retrieving all correspondence authored by any individual who resided in a specific province during a predefined ten-year span. These qualitative successes underscore that the true value of entity linking in cultural heritage lies not merely in tagging text, but in generating a cohesive, structured, and inherently interoperable historical environment.

5. Conclusion and Future Directions

5.1 Summary of Methodological Contributions

This comprehensive analysis has demonstrated that the persistent challenge of fragmentation within cultural heritage archives can be effectively addressed through the rigorous application of entity linking and semantic grounding methodologies. By systematically extracting named entities from unstructured, historical texts and resolving their identities against standardized knowledge bases, institutions can transcend the limitations of localized, heterogeneous cataloging systems. The detailed architectural pipeline presented in this study, which encompasses specialized historical language modeling for recognition, context-aware dense embeddings for disambiguation, and formal ontology alignment for semantic structuring, provides a robust blueprint for digital humanities integration. The experimental evaluations confirmed that while historical texts present unique complexities such as severe orthographic variation and entity ambiguity, modern contextual approaches significantly outperform traditional string-matching techniques, particularly in the highly accurate integration of geographic and organizational data. Ultimately, semantic grounding transforms static digital artifacts into a dynamic, interconnected web of historical knowledge, facilitating seamless interoperability across the global GLAM sector.

5.2 Practical Implications and Limitations

The practical implications of implementing these methodologies are substantial for both archival institutions and historical researchers. For institutions, automated record integration reduces the massive manual labor traditionally required to harmonize diverse collections, allowing them to participate in the broader Linked Open Data ecosystem with greater efficiency. For researchers, the transition from document-level retrieval to entity-centric knowledge graphs enables unprecedented macro-level analyses, uncovering hidden sociopolitical networks and historical trends that span multiple archives and geographic boundaries. However, the study also acknowledges several significant limitations. The performance of entity linking systems remains heavily dependent on the comprehensiveness of the target knowledge bases, which often exhibit historical and western-centric biases, resulting in the marginalization of lesser-known regional entities [16]. Furthermore, the computational resources required to train specialized historical language models and maintain vast ontological graphs pose a barrier to entry for smaller, underfunded archives. Future research must therefore focus on developing highly efficient, low-resource disambiguation models and expanding historical knowledge bases to ensure more equitable and globally comprehensive record integration. Ongoing interdisciplinary collaboration between computer scientists and domain experts will be essential to refining these semantic technologies and fully realizing the potential of integrated digital heritage.

References

1. Ardi, H.; Rusmin, P.H.; Abrari, M.G. Circular Arc Based Path Smoothing for Precise Forklift Navigation and Orientation Alignment. In Proceedings of the 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI), Jakarta, Indonesia, 21–21 January 2025; pp. 176–181.

2. Xiong, Y.; Varadarajan, B.; Wu, L.; Xiang, X.; Xiao, F.; Zhu, C.; Dai, X.; Wang, D.; Sun, F.; Iandola, F.; et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16111–16121.

3. Morbidini, F.; Samanta, A.; Maucieri, C.; Karydis, K.; Mauk, P.A.; Skaggs, T.H.; Scudiero, E. Robotic mapping of soil volumetric water content with geospatial soil apparent electrical conductivity in micro-irrigated citrus orchards in California. Comput. Electron. Agric. 2026, 245, 111540.

4. Zhang, Z.; Xiao, B. The Influence of Cargo Moving and Sliding Mode Control Strategy for Forklift. IEEE Access 2020, 8, 16637–16646.

5. Kita, N.; Kato, T. Image Measurement Method for Automatic Insertion of Forks into Inclined Pallet. In Proceedings of the 2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, Singapore, 11–13 December 2022; pp. 441–448.

6. Rogachev, G.; Patkin, M.; Rogachev, N. Fuzzy Optimization in the Problem of Forklifts Path Planning. In Proceedings of the 2018 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), Vladivostok, Russia, 3–4 October 2018; pp. 1–5.

7. Giglio, D. Task scheduling for multiple forklift AGVs in distribution warehouses. In Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA), Barcelona, Spain, 16–19 September 2014; pp. 1–6.

8. Zheng, F.; Lu, Y.; Fang, M.; Li, L. Conflict Model and Time Window Based Multi AGF Path Finding and Coordination. In Proceedings of the 2023 IEEE International Conference on Real-time Computing and Robotics (RCAR), Datong, China, 17–20 July 2023; pp. 408–413.

9. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.

10. Lv, S.; Wang, S.; Liu, J. RGB-Image-Based Forklift Pallet Recognition and Localization. In Proceedings of the 2023 6th International Conference on Software Engineering and Computer Science (CSECS), Chengdu, China, 22–24 December 2023; pp. 1–5.

11. Chen, G.; Peng, R.; Wang, Z.; Zhao, W. Pallet Recognition and Localization Method for Vision Guided Forklift. In Proceedings of the 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China, 21–23 September 2012; pp. 1–4, ISSN 2161-9654.

12. Raharja, K.A.W.; Rusmin, P.H. Detection and Measurement of Object Distances Using the Combination Function of 2D LiDAR Sensor and Camera. In Proceedings of the 2023 International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia, 10–11 October 2023; pp. 1–5.

13. Tiozzo Fasiolo, D.; Scalera, L.; Maset, E.; Gasparetto, A. The FriûlBot dataset: Experimental validation of an autonomous ground robot for vineyard 3D mapping. Robot. Auton. Syst. 2025, 197, 105283.

14. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. arXiv 2025, arXiv:2509.25164.

15. Rösmann, C.; Hoffmann, F.; Bertram, T. Integrated online trajectory planning and optimization in distinctive topologies. Robot. Auton. Syst. 2017, 88, 142–153.

16. Bhat, A.; Kai, N.; Suzuki, T.; Shiroshima, T.; Yoshida, H. Fake Obstacle Generation-based Pallet Alignment for Forklifts. In Proceedings of the 2024 16th International Conference on Computer and Automation Engineering (ICCAE), Melbourne, Australia, 14–16 March 2024; pp. 676–680.