

# A METHODOLOGICAL APPROACH TO ASSET INFORMATION MANAGEMENT VIA KNOWLEDGE GRAPHS AND LARGE LANGUAGE MODELS

Marta Boscariol<sup>1,\*</sup>, Silvia Meschini<sup>2</sup> and Lavinia Chiara Tagliabue<sup>2</sup>

<sup>1</sup> Department of Management, Università degli Studi di Torino, 10134 Turin, IT

<sup>2</sup> Department of Computer Science, Università degli Studi di Torino, 10149 Turin, IT

\* marta.boscariol@unito.it

## Abstract

Tackling the need of large organizations for a proactive Asset Information Management (AIM) System, a methodological approach to knowledge management applied to built assets portfolios is proposed. It aims at synergically leveraging Knowledge Graphs (KGs) and Artificial Intelligence (AI) technologies to enable analytics on input data. In the theorized pipeline Large Language Models (LLMs) are meant to be used both in the graph creation phase, extracting data from unstructured sources and organizing them according to domain ontologies, as tested on a use-case sample, and in the knowledge extraction phase via queries.

## Introduction

Asset Management (AM) is defined by international standardization as “coordinated activity of an organization to realize value from assets”, where an asset is identified as an “item, thing or entity that has potential or actual value to an organization” (ISO 55000, 2014). It can be applied to the construction sector in terms of balancing of financial aspects, extraordinary interventions, operations and maintenance, in order to provide the most cost-effective, sustainable, well-engineered and efficient solution. The operational phase of buildings and infrastructures (ISO 19650-1, 2019), represents that time span when a continuous stream of minor interventions take place and the asset itself both delivers value to the organization and represents a source of costs. Asset owners with large portfolios typically manage numerous facilities, having the majority of them in the operational phase (UK BIM Framework, 2021).

An AMS, “whose function is to establish the asset management policy and asset management objectives” (ISO 55000, 2014), bases its performance on data exploitation, since assets can be considered data-rich environments, or recollections of a multitude of data and documents depicting information about each building’s lifecycle (LC) step. From a data management perspective, future availability and accessibility of information is required, for at least as long as the life of the asset, jointly with pre-agreed Master Data Management (MDM) (ISO 8000-100, 2016) procedures internal to the organization, to support the effective management of data quality and deliver benefits across various LC phases.

Provided the output of this investigation is suggesting an Asset Information Management (AIM) approach to support the managerial processes, information ought to be

regarded as a precious asset itself, whose structure and consistency provide measurable performance indicators and insights for decision-making purposes (KPMG, 2021). The main challenges associated with managing built asset data are related to their heterogeneous and siloed nature, being their structure and storage format quite often not regulated and then handled by multiple users. These dynamics determine a loss of value related to the information managed per each step of the process (Eastman, 2011), proving to be obstacles to the extraction of valuable and complete information in a timely manner. Managers and other stakeholders involved should be enabled to store, access and query information regarding their portfolio, ranging from documents to spatial information, 3D models or activity-related materials, no longer dealing with inaccurate, incomplete or even undiscoverable pieces of documentation. Inappropriate creation and maintenance of data sources, as well as interoperability issues foster higher management costs and a lack of trust in the final user, who’s trying to gain a deeper insight. Recently, with the advent of Big Data (Ajah and Nweke, 2019), innovative technology solutions enable organizations to improve operational efficiency (Gomolluch, 2023), strengthen stakeholder relations, reduce costs and provide demonstrable legal compliance. Big Data Analytics, as a topic, has found some application also in the built environment to foster decision making (Farghaly et al., 2017, Demirdöğen et al., 2023). Limiting their implementation with BIM tools might not be the most efficient way to manage such intense streams of data, due to inadequate storage capabilities of those technologies. The goal of the current research is to suggest a holistic (Gomolluch, 2023) practical approach to knowledge management, applied to extended built assets portfolio, such as buildings and private or public infrastructures, through the adoption of Knowledge Graphs (KGs), as well as Large Language Models (LLMs) automating otherwise lengthy manual workloads. This proposition stems from a real use-case scenario, which is the extended built assets portfolio managed by the University of Turin. In this public institution an AMS (Di Giuda et al., 2023) is currently under development, dealing with the above mentioned limitations related to siloed documental resources and data quality. The proposed methodology is willing to enable the extraction of knowledge from heterogeneous data sources and their proactive interrogability, therefore improving the understanding of how real estate assets are used, along with decision-making at every stage of the asset’s LC.

## Literature background

Scientific literature in the field of AM in the built environment lacks a unitary vision and covers the topic of Information Management through sectorial applications. AM is often investigated employing BIM (Farghaly et al., 2017, Moretti et al., 2022) or through Digital Building Logbook as documental repositories of a building's lifecycle data, with a focus on sustainability and energy efficiency (Signorini et al., 2023, Malinovec Puček et al., 2023). When it comes to applying KGs and Linked Data to the built environment (Pauwels et al., 2022), it is often related to the definition of ontologies fit for specific use cases and their application to BIM models: sensors and mechanical systems (Quinn et al., 2020, Eleftheriou et al., 2022); building components (Sobhkhiz and El-Diraby, 2023); operation and maintenance (Xie et al., 2022); thermal comfort and energy analysis (Esnaola-Gonzalez et al., 2022); construction phase (Farghaly et al., 2021), even considering the integration of open source BIM models (IFC based) (Elshani et al., 2022). Others investigated link validation in ICDDs (Information Containers for Linked Document Delivery), meant for linking information in documents with Resource Description Framework (RDF) data (Hagedorn, Liu, et al., 2023, Hagedorn, Pauwels, et al., 2023), or exploited Shapes Constraint Language (SHACL) applied to distributed Linked Building Data for building model checking and constraint validation (Nisbet et al., 2023). Differently, KGs as a representation of Digital Twins has been studied in other sectors, mainly in industry 4.0 (Yahya et al., 2021, Tamašauskaitė and Groth, 2023), while text-based knowledge discovery techniques (i.e., Text Mining, Text processing, and Natural Language Processing) are hardly applied to the building sector. Natural Language Processing (NLP), which is branch of Artificial Intelligence (AI) has found application in text analysis and compliance checking related to contracts and legislation (Locatelli et al., 2023). In the field of AI and LLMs, algorithms are being developed to deal with the extraction of content from multiple kind of sources, both structured databases (Sequeda et al., 2023) and text-like documents, proving promising in assisted and semi-automated KG construction pipelines (Zhu et al., 2023).

## Research context and aim

In compliance with studies stating that organizations can benefit from implementing AMS conforming to the ISO 55000 normative series (Almeida et al., 2022, Maletič et al., 2023), the proposition of a practical AIM approach should consider some recurrent characteristics in big organizations managerial praxis:

- Widespread organizations usually store large volumes of data and documentation, heterogeneous in types and formats, in centralized repositories. So called data lakes ingest all types of data from any source, ranging from structured (database tables, .xlsx sheets) to semi-structured (XML, webpages) to unstructured (images, audio files, PDFs), in their original raw form. Due to data lakes scalable architecture, a variety of

applications can be compatible (Sawadogo and Darmont, 2021), from big data analytics to machine learning (ML) or predictive analytics, therefore greater value can be extracted from stored data.

- Inside an organization, analogous or related tasks are not always systematically carried out using harmonized terminology or producing consistent documentation, resulting in misalignments in both structure and content of files. Semantic enrichment of raw data would ensure clear interpretation and understanding, enabling the adoption of a graph structure (Pan et al., 2024) to interconnect individual records in a network and overcome inconsistencies.
- Structuring a uniform framework of machine-readable data generates benefits (Gomolluch, 2023) both in the short term, providing a Knowledge Base (KB) agile to interrogate, and in the long term, posing the basis for further ML implementations, which might include prediction algorithms and automatizations, optimizing managerial processes.

The theorized approach aims at multiplying the advantages provided by AI-powered technologies in Information Management, especially in decision-making processes at a managerial level. In a dual sided technological pipeline, from one side there's an AI system automatically extracting information from document sources and structuring a KB as a graph, while on the other side there's a human user interrogating another AI agent, capable of performing neural-symbolic reasoning (Zhang et al., 2021). This system, enriched by structured knowledge and logic-based inference from ontologies, based on the relevant facts retrieved from the KB, should enable informed decision making.

## Methodology

As previously stated, the core idea of this research project deals with the development of a methodology to gather all sources of knowledge regarding multiple built assets in a unique interrogable environment, structured as a KG. The methodology (*Figure 1*) is composed of two parts: an *automated text-to-KG pipeline* and a *chatbot for human-machine interaction*. The former is responsible for the extraction of every bit of information existing in the organization's documents corpora, for their labeling according to pre-established sets of terms and for their interrelation to generate a KB with a graph structure. The latter assists users in finding the information they require from the KG. It should act as user-friendly interface and extract information from the KG through an NLP agent, which transforms natural language queries into query language (e.g. SPARQL), drawing inference and returning search results in human readable form. Its purpose is to perform those analytics typically carried out manually and visualizing them through graphical forms and set Key Performance Indicators (KPIs) into the AMS digital boardroom, although its proposition will be subject to future developments.

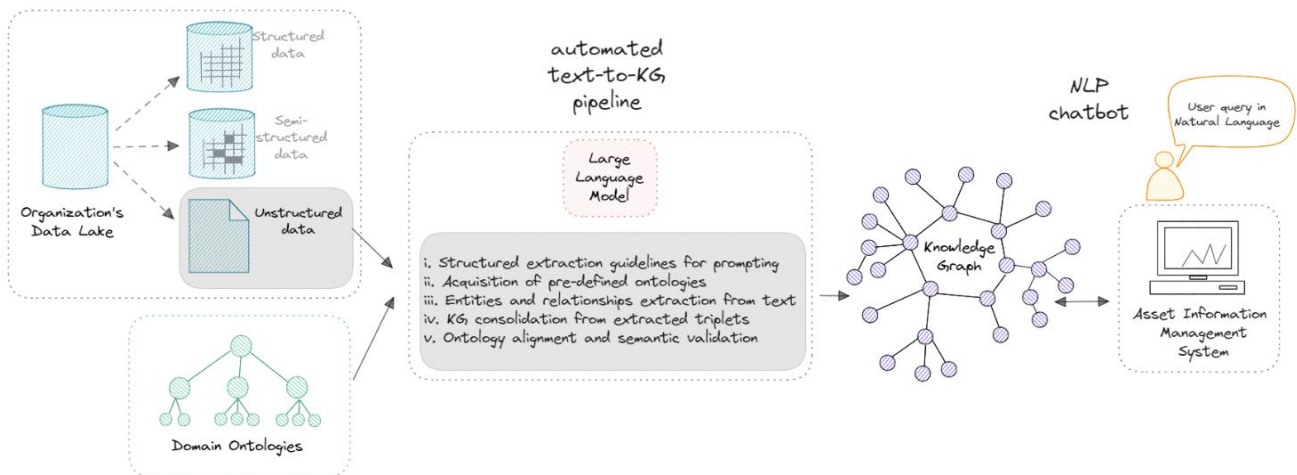


Figure 1. Theorization of an Asset Information Management approach

Prior to the tools development stage, some considerations ought to be made, regarding the intended scope of an AIM digital tool, according to the organization's administrative activities related to its use. For this purpose, a team of experts should identify one or more Domain Ontologies relevant to AIM scopes, creating those that match the organization's requirements, formulated in terms of competency questions (Keet, 2020). In addition to specifically created ontologies, existing and validated ones could be reused.

The definition of the methods to be applied in order to extract a KG from the enterprise Data Lake, based upon heterogeneous data sources, both structured and non-structured, strongly depends on the use case scenario and it has been chosen to deal with text documents. The suggested methodology aims at integrating what seems to be two distinct technologies, LLM and KGs, which though aim both at capturing a detailed representation of a portion of the world. Their different knowledge representation formats complement each other, overcoming individual limitations. A so called *text-to-KG* pipeline will be implemented, exploiting literature-tested solutions, and it involves:

- a Large Language Model (LLM), which acquires world knowledge by pre-training on massive text corpora and is capable of understanding and transforming human language, although with severe likelihood of hallucinations or logical leaps, showing limitations in domain-specific performance, trustworthiness, controllable generation or content quality assessment.
- KGs, which directly store factual knowledge as interconnected networks via relationships of real-world entities and provide structured, symbolic knowledge and deterministic logic, but struggle at handling nuanced language. They ground LLMs in factual knowledge, ensuring truthful text generation, while, on the other hand, LLMs capabilities in text processing and fact extraction mitigate their rigidity.
- Ontologies, which represent structured formal knowledge, including commonsense knowledge

lacking in LLMs, and are used to define concepts, properties, relationships, constraints, and axioms in a machine-readable format, using standards like RDF and Web Ontology Language (OWL). They play the role of “organizing principle”, as they provide agreed sets of hierarchies and terminologies, harmonizing the content handled and permitting entity disambiguation, as well as describing with the required level of detail the domain of interest. Ontologies act as a bridge between the unstructured textual knowledge of LLMs and the structured world knowledge represented in KGs, improving symbol grounding and enabling formal logic reasoning (Keet, 2020), such as inference through querying protocols.

The exploitability of the LLM-driven *text-to-KG* pipeline is strictly dependent on the domain ontology, as, in the first place, they will help the AI algorithm extract the correct entities and relationships. Although their outlining is an activity that requires time and the involvement of many stakeholders, it can be seen as a more formal task, upon which the technological pipeline can be built, as in the two steps described onwards.

#### Domain Ontologies development: the organizing principle

AM information ought to be systematized with the help of domain-specific ontologies, through classes, their properties and relations. Since “an ontology is a formal, explicit specification of a shared conceptualization” (Studer et al., 1998), the task of ontology engineering has to be demanded to a group of domain experts, jointly with professionals competent in RDF and OWL ontologies matter. Considering that concepts to be harmonized are already in use among the organization’s practitioners, a bottom-up approach is suggested (Keet, 2020). Core concepts and relationships, as they tend to appear in datasets, are to be stated and formalized in the main ontology first, along with formal definitions via ontological axioms. Local ontologies and their alignment will be implemented only after the whole framework and its applicability have been validated, resulting in an incremental approach.

It is to be remarked that the core aim of the methodology is to create KGs conforming to a specific ontology, in order to guide accordingly to the schema all queries posed to the resulting graph, as well as to process a large number of documents through multiple prompts, due to computational size limitations.

### LLMs for Knowledge Graph generation from text: the technological pipeline

Given recent advancements of LLMs and foundation models, they could be leveraged to alleviate time-consuming tasks, dealing with a great amount of unstructured files stored in the organization's repositories. To better understand how the *text-to-KG* pipeline could prove successful, a disaggregation of its procedure (Figure 1) is therefore outlined:

- i. Creation of Structured Extraction Guidelines, whose scope it's the definition of the types of entities, properties and relationships the LLM needs to recognize and extract, referring to the formal content of the use-case ontology. Guidelines will be supported with input samples from the target domain and examples of the expected output structure in ontology language, to teach the LLM how to correctly encode human logic and to avoid inconsistencies in information extraction.
- ii. Use of a LLM to transform unstructured text to KG in compliance with a pre-defined ontology. This task could be accomplished using either a LLM pre-trained with standard ontologies, a LLM fine-tuned with a custom ontology, a LLM prompted with a custom ontology, or a hybrid of a fine-tuned (with a custom ontology) and pre-trained (with a standard ontology) LLM. A comparative analysis investigating the most fit-for-purpose model should be conducted, evaluating benefits and limitations of each of the four plausible approaches outlined, taking into account token costs, response times, training and tuning datasets required. A hybrid approach is considered more suitable for high-quality KG construction (Wang et al., 2023), relatively to its accurate extraction of domain ontology components.
- iii. Extraction of entities and relationships from large volumes of text sources via LLM, automating the process of KG extraction through text chunking, extraction of triplets, their normalization and enrichment. Subsequently text mining techniques and NLP algorithms (Van Assche et al., 2023, Zhao et al., 2023) operate on entities and their relationships.
- iv. KG consolidation, merging nodes and edges extracted from all sources into a unified set and enhancing them with metadata and provenance information, to allow topological analysis on communities and centralities. The resulting KG is ready to be uploaded to a storage database, becoming a unitary source of data, over which complex analytics and queries could be carried out. Further integrations of departmental KGs, deriving from other domain ontologies, are allowed through an incremental approach.

- v. Ontology alignment or ontology matching (OM) takes ontologies as input and determines as output a set of correspondences between semantically related entities, dealing with semantic heterogeneity problems (Li et al., 2019) and enabling navigation over KGs.

A few-shot prompting framework has been tested, to guide LLMs in performing NLP tasks, finalized at KG creation, exploiting GPT-3.5 (gpt-3.5-turbo), despite it being a closed-source model. The authors tried to define an ad-hoc chaining prompts to guide the LLM through a series of smaller, more manageable tasks to ultimately achieve a complex goal, which is the conversion of an unstructured paragraph into a structured graph, where each node corresponds to an entity, and the interconnecting edges depict the relationships linking them, matching those defined in a sample ontology. An exploratory trial-and-error experiment has been conducted, feeding a use-case textual source, in which university spaces are listed and assigned to organizational units and their occupants, and a reduced version of the RealEstateCore (REC) Ontology. The prompts engineered go through the following steps:

1. Task definition: the LLM is given a role as expert and is told to perform information extraction from textual documents, converting it into triplets made as: Source\_Node, Relationship, Target\_Node. Additional instructions are here provided, concerning how to maintain consistency in entities and handle numerical data, property formats, naming conventions.
2. Ontology ingestion: the LLM is fed with an ontology in RDF compatible .ttl format, being instructed on which entities and relationships to look for in the following tasks. Due to tokens limitations, a small ontology, composed only of use-case related entities and relationships, has been exploited.
3. Unstructured paragraph input: the LLM is given as input the text containing data aimed at being converted to KG. It is preferably accompanied by an example of how information should be correctly extracted, stating a small paragraph and the desirably outputted list of triples that could be identified in that paragraph.
4. RDF serialization: the LLM is asked to serialize the list of extracted entities and relationships according to RDF format, in this case in .ttl, enabling compatibility with graph databases.

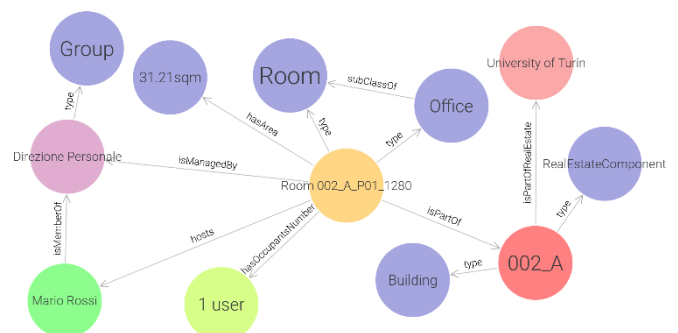


Figure 2. Visualization of nodes connected to a room entity in Ontotext GraphDB.

The resulting file was imported into an open source graph store, namely Ontotext GraphDB, where triplets consistency with the original NL source has been verified. As an example, nodes and relationships connected the one of the room entities extracted from text has been visually displayed as a graph (*Figure 2*).

## Discussion

### Theoretical contributions

Literature demonstrates that a single technology platform can be exploited to create accurate and transparent records of organizations' reference data (including buildings, assets, people, and contracts) and it can be integrated in internal business processes (e.g. contract and project management, strategic planning, environmental reporting, preventative and reactive maintenance) (Hanley and Brake, 2016, Geisler et al., 2021). However, a strategic approach to implementation, which must align to business process change and strong data governance, is required.

From an Information Management perspective, the process as a whole should be applied proportionally to the scale and complexity of the organization and of its asset portfolio, reflecting concepts and principles from ISO 19650-1 (UK BIM Framework, 2021). Accordingly, agreed methods and procedures should be internally standardized, to ensure the production of compliant set of data and information deliverables. Agreed schemata would ensure compatibility when federating data produced by different internal departments, which can then be used and maintained during the asset's operational life for analysis, reporting and decision-making.

### Practical implications and limitations

The proposed methodological approach to knowledge management, applied to built assets' portfolios, aims to leverage a KG structure and AI technologies to enable analytics on input data for a proactive Asset Information Management System in large organizations. From an enterprise management perspective, a KG approach could solve data analytics challenges related to data quality or complex properties, exploiting intrinsic characteristics of Data Lakes storing the sources of knowledge related to accessibility, enabling scalability through a domain oriented structure and integration across all domains of the organization's network.

The main advantages deriving from the exploitation of semantic web technologies, in particular of KGs and ontologies, can be recognized in the automatic interrelation of information content, without having a human operator search for a specific document in a multitude of folders, and in its being univocally classified, reducing data creators' margins of personalization. With an incremental approach, from a single KB might stem several use-case applications, as multiple domain-specific ontologies can coexist in the same KG, putting in place an iterative construction of the KG introducing "just enough" semantics as each use case demands.

Integrating ontologies enable formal logical inference, complementing the inductive reasoning of LLMs with description logics and enhancing their generalization capabilities, as symbol grounding explicitly links language to formal conceptual representations. Moreover, a coexistence of standard and custom ontologies is feasible, as well as both independent and cross equivalence queries across domains. Additional benefits could be given by the adoption of FAIR (Findable, Accessible, Interoperable, Reusable) Data Principles, as they provide a data management framework, enabling findability of data and metadata through persistent digital object identifier (DOI); accessibility to trusted repositories; interoperability of knowledge representation language and dictionaries; reusability of data, thanks to accurate provenance information and clear usage license (Martorana et al., 2022).

Other potential advantages of using an automated methodology for KG construction from unstructured data at scale resides on saving a great amount of hours otherwise required for manual reading, labeling and linking of documents, as for inconsistencies and redundancies checking in their content. LLM integration further differentiates this methodology from traditional KG creation, accomplishing the main aim of the final KG, namely inference and query answering according to explicitly defined schema and properties, along with ensured high volume scalability, easier model retraining, debugging and dataset or prompt fixings.

According to the authors observations during the testing phase, the accuracy of the generated knowledge graph clearly relies on the underlying model leveraged, as the extracted information might be affected by biases and hallucinations, given by the LLM's ability to correctly interpret its tasks. Prompt development involved an interactive process of writing, testing, and refinement, based on feedback and experimentation. Tailoring the prompt to use-case specific application required giving the LLMs sufficient context to generate the desired output. This includes writing clear step-by-step instructions, as detailed as possible about the outcome format. Instead than zero-shot prompting (when the LLM relies solely on its pre-existing knowledge and on the instructions in the prompt to perform the task), adopting few-shot learning enabled the LLM to recognize patterns from little examples provided and to generalize to unseen task variations, also avoiding time-consuming data labeling. It emerged how the model tends to assign all relationships extracted from text to the ontology proposed, even those not standardized by it, and it struggles to infer transitive and inverse relationships from the ontology, if not explicitly declared in the example as well. Other limitations encountered testing this prompting framework rely on computational limits, which, in the case of GPT-3.5, are set to 4'096 tokens per prompt. Therefore textual sources to be analyzed must be pre-processed and split into smaller size compliant chunks, taking care of adequately preventing the loss of context

and the separation of related entities (Carta et al., 2023). In order to ensure scalability for larger datasets, an algorithm devoted to chunking text and automatically and iteratively prompting the LLM should be implemented.

The proposed methodology presents intrinsic limitations, related to the time and expertise required to develop a proper domain ontology, as well as to the implementation and testing phase of the technological pipeline. It will be highly dependent on the application use-case, mainly caused by the uncertainty generated by the sample size and to the unreliability of its data sources. Due to the lack of prior research studies on this precise topic, it will be difficult to compare results and assess their validity, requiring the establishment of validation methods.

### Future directions

The potential offered by this approach in the field of Knowledge Representation and Reasoning (KRR) is manifest, since KGs have emerged as key technologies for representing knowledge in a variety of domains and supporting intelligent applications, such as chatbots, question answering systems, and recommendation systems (Hu et al., 2023). An NLP agent devoted to answering a natural language question, by reasoning over facts in a KB, could potentially deal with text-to-SPARQL tasks and multi-hop reasoning, appropriately querying the organization's KG on user's behalf. These technologies should provide also faster access to continuously-updating data like sensor readings, getting closer to real-time analytics.

Other future developments might consider multiple domains integration, differential privacy policies, implementing user-based restrictions on visualization, as well as the integration of ML algorithms for predictions. Reinforcement learning from human feedback and automatic updates to the KG as new documents are ingested, will make the methodology more versatile.

As previously stated, the research project will leverage a practical use-case methodological approach, not only suggesting a plausible use of digital technologies to tackle AM processes inside the University of Turin, but also aiming at providing a scalable, generalizable and reusable set of tools, which might solve similar knowledge-related problems in other domains. Relatively to the applicative use-case, textual documents containing information regarding ownership and occupancy of the building are employed. The implementation of a *text-to-KG* will extract and interconnect factual knowledge, applying a customized ontology as “organizing principle”. The resulting KG's consistency will be evaluated through recall and precision metrics based on confusion matrix entries, obtained from manual evaluation campaigns.

Aligning with past projects at the University of Turin, where Business intelligence (BI) tools have been implemented as means of data-driven decision making (Di Giuda et al., 2023), an integration of AI and BI would be desirable, to drive faster, automated and predictive insights. The use of AI powered analysis tool could

overcome traditional BI tools limitations regarding analysis at a more granular level, benefitting the end user in terms of automating data cleaning to ensure accuracy in the procedure, integrating multidimensional analysis as well, to detect clusters or similarities across hierarchies, attributes and metadata stored in the graph database.

### Conclusions

The paper tackled the proposition of a methodological approach to AIM, leveraging technological pipelines whose task is the automation of truthful knowledge extraction from a large organization's documental corpora, to enable informed decisional processes. The pivotal element is the human user, as it determines specific information needs guiding the ontology creation, as well as practical decision-making goals and problem-solving context defined by the organization it represents. KGs provide structured knowledge by encoding concepts and relationships in machine-readable ontologies, therefore they can deliver accurate, safe and responsible AI applications, as they drive reasoning and semantics. However, the creation of an enterprise KGs manually might involve gigantic and cost-inefficient workload, hence the automation of ontology extraction from unstructured corpora through LLMs could be a viable solution. As it has been outlined, the incorporation of ontologies in the pipeline can significantly strengthen LLMs reasoning capabilities and structure accurate KGs, overcoming present challenges in reasoning and symbol grounding, granting a significant increase in the value of an organization's knowledge graph. The authors tested a semi-automated approach to KG construction for the AIM domain, leveraging few-shot LLM prompting with GPT-3.5 model. The chaining prompting technique dealt with triplets extraction from unstructured sources, semantically aligning them with a pre-processed ontology acting as T-Box. The resulting KG showed GPT-3.5 proficiency in NLP tasks applied to closed-domain settings, taking advantage from clearly stated tasks explanation, although hallucinating ontology classes. Since being tested only on a small dataset, the authors intend to conduct more in-depth experimentation, to properly evaluate reliability and fitness to the intended purpose of the approach, qualifying replicability and generalizability, also evaluating alternative LLMs. Future developments will likely integrate other AI systems on the inference side of the pipeline, capable of dealing with continuously changing data and forecasting plausible event scenarios.

### Acknowledgements

It is to be remarked that sensitive data contained in sources provided by the University of Turin have been anonymized prior to AI tools exploitation. The authors also want to thank Dr. Mirko Locatelli for sharing his knowledge on NLP and for engaging in enriching conversations on the topic.

## References

- Ajah, I. A. & Nweke, H. F. (2019) Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications. *Big Data and Cognitive Computing*. Multidisciplinary Digital Publishing Institute, 3(2), p. 32.
- Almeida, N., Trindade, M., Komljenovic, D. & Finger, M. (2022) A conceptual construct on value for infrastructure asset management. *Utilities Policy*, 75, April, p. 101354.
- Buchholz, M. & Lützkendorf, T. (2023) European building passports: developments, challenges and future roles. *Ubiquity Press*, 4(1), pp. 902–919.
- Carta, S., Giuliani, A., Piano, L., Podda, A. S., Pompianu, L. & Tiddia, S. G. (2023) Iterative Zero-Shot LLM Prompting for Knowledge Graph Construction. *arXiv*.
- Demirdöğen, G., Işık, Z. & Arayıcı, Y. (2023) BIM-based big data analytic system for healthcare facility management. *Journal of Building Engineering*, 64, April, p. 105713.
- Di Giuda, G. M., Meschini, S., Accardo, D., Gasbarri, P., Tagliabue, L. C. & Scomparin, L. (2023) BIM-GIS and BI integration for facility and occupancy management of university assets: the UniTo pilot case. In: Florence, Italy.
- Eastman, C. M. (ed.) (2011) *BIM handbook: a guide to building information modeling for owners, managers, designers, engineers and contractors*. 2nd ed, Hoboken, NJ: Wiley.
- Eleftheriou, O., Dimara, A., Kotis, K. & Anagnostopoulos, C.-N. (2022) Saving Energy with Comfort: A Semantic Digital Twin Approach for Smart Buildings. In *Proceedings of the Third International Workshop on Semantic Digital Twins*. Hersonissos, Greece: CEUR (CEUR Workshop Proceedings).
- Elshani, D., Wortmann, T. & Staab, S. (2022) Towards Better Co-Design with Disciplinary Ontologies: Review and Evaluation of Data Interoperability in the AEC Industry. In *Proceedings of the 10th Linked Data in Architecture and Construction Workshop*. Hersonissos, Greece: CEUR (CEUR Workshop Proceedings), pp. 43–52.
- Esnaola-Gonzalez, I., Bermúdez, J. & Aceta, C. (2022) An Ontology-Driven Approach to Support Data Analysts with Thermal Comfort Problems in the Built Environment. In *Proceedings of the 10th Linked Data in Architecture and Construction Workshop*. Hersonissos, Greece: CEUR (CEUR Workshop Proceedings), pp. 8–19.
- Farghaly, K., Abanda, H., Vidalakis, C. & Wood, G. (2017) BIM Big Data System Architecture for Asset Management: A Conceptual Framework. In *Lean and Computing in Construction Congress - Volume 1: Proceedings of the Joint Conference on Computing in Construction*. Heraklion, Crete, Greece: Heriot-Watt University, pp. 289–296.
- Farghaly, K., Soman, R. K. & Whyte, J. (2021) Visualizing real-time information through a construction production control room. In *European Council on Computing in Construction (Computing in Construction)*, pp. 415–422.
- Geisler, S., Vidal, M.-E., Cappiello, C., Lóscio, B. F., Gal, A., Jarke, M., Lenzerini, M., Missier, P., Otto, B., Paja, E., Pernici, B. & Rehof, J. (2021) Knowledge-Driven Data Ecosystems Toward Data Transparency. *Journal of Data and Information Quality*, 14(1), p. 3:1-3:12.
- Gomolluch, A. B., M. Block, D. Gogolin, S. (2023) Potential of holistic asset information management. In *Life-Cycle of Structures and Infrastructure Systems*. CRC Press.
- Hagedorn, P., Liu, L., König, M., Hajdin, R., Blumenfeld, T., Stöckner, M., Billmaier, M., Grossauer, K. & Gavin, K. (2023) BIM-Enabled Infrastructure Asset Management Using Information Containers and Semantic Web. *Journal of Computing in Civil Engineering*. American Society of Civil Engineers, 37(1), p. 04022041.
- Hagedorn, P., Pauwels, P. & König, M. (2023) Semantic rule checking of cross-domain building data in information containers for linked document delivery using the shapes constraint language. *Automation in Construction*, 156, December, p. 105106.
- Hanley, B. P. & Brake, D. J. (2016) Putting asset data at the heart of organisational decision-making using an Integrated Workplace Management System. In *Asset Management Conference (AM 2016)*, pp. 1–7.
- Hu, N., Wu, Y., Qi, G., Min, D., Chen, J., Pan, J. Z. & Ali, Z. (2023) An empirical study of pre-trained language models in simple knowledge graph question answering. *World Wide Web*, 26(5), pp. 2855–2886.
- ISO 8000-100 (2016). Switzerland: ISO. ISO/TC 184, Automation systems and integration, Subcommittee SC 4, Industrial data.
- ISO 19650-1 (2019). London: British Standards Institution.
- ISO 55000 (2014). Switzerland: ISO. ISO/PC 251 Asset management.
- Keet, M. (2020) *An Introduction to Ontology Engineering*.
- KPMG (2021) *The value of Information Management in the construction and infrastructure sector*. A report commissioned by the University of Cambridge's Centre for Digital Built Britain (CDBB), p. 104.
- Li, H., Dragisic, Z., Faria, D., Ivanova, V., Jiménez-Ruiz, E., Lambrix, P. & Pesquita, C. (2019) User validation

- in ontology alignment: functional assessment and impact. *The Knowledge Engineering Review*. Cambridge University Press, 34, January, p. e15.
- Locatelli, M., Pattini, G., Pellegrini, L., Meschini, S. & Accardo, D. (2023) Fostering the consensus: a BERT-based Multi-label Text Classifier to support agreement in public design call for tenders. *TEMA: Technologies Engineering Materials Architecture*, 9(1), pp. 62–73.
- Maletič, D., Grabowska, M. & Maletič, M. (2023) Drivers and Barriers of Digital Transformation in Asset Management. *Management and Production Engineering Review*, 14(1), pp. 118–126.
- Malinovec Puček, M., Khoja, A., Bazzan, E. & Gyuris, P. (2023) A Data Structure for Digital Building Logbooks: Achieving Energy Efficiency, Sustainability, and Smartness in Buildings across the EU. *Buildings*. Multidisciplinary Digital Publishing Institute, 13(4), p. 1082.
- Martorana, M., Kuhn, T., Siebes, R. & Ossenbruggen, J. van (2022) Aligning restricted access data with FAIR: a systematic review. *PeerJ Computer Science*. PeerJ Inc., 8, July, p. e1038.
- Moretti, N., Xie, X., Garcia, J. M., Chang, J. & Parlikad, A. K. (2022) Built environment data modelling: a review of current approaches and standards supporting Asset Management. *IFAC-PapersOnLine*. (5th IFAC Workshop on Advanced Maintenance Engineering, Services and Technologies AMEST 2022), 55(19), pp. 229–234.
- Nisbet, N., Zhang, Z. & Ma, L. (2023) Automated generation of SPARQL queries from semantic markup. In. *European Council on Computing in Construction (Computing in Construction)*, pp. 0–0.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. & Wu, X. (2024) Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* pp. 1–20.
- Pauwels, P., Costin, A. & Rasmussen, M. H. (2022) Knowledge Graphs and Linked Data for the Built Environment. In *Industry 4.0 for the Built Environment: Methodologies, Technologies and Skills*. Cham: Springer International Publishing (Structural Integrity), pp. 157–183.
- Quinn, C., Shabestari, A. Z., Misic, T., Gilani, S., Litoiu, M. & McArthur, J. J. (2020) Building automation system - BIM integration using a linked data structure. *Automation in Construction*, 118, October, p. 103257.
- Sawadogo, P. & Darmont, J. (2021) On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56(1), pp. 97–120.
- Sequeda, J., Allemang, D. & Jacob, B. (2023) A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. arXiv.
- Signorini, M., Moretti, N., Merino, J., Daniotti, B. & Parlikad, A. (2023) Digital-twin based data modelling for digital building Logbook implementation. *IET Digital Library*, January, pp. 130–137.
- Sobhkhiz, S. & El-Diraby, T. (2023) Dynamic integration of unstructured data with BIM using a no-model approach based on machine learning and concept networks. *Automation in Construction*, 150, June, p. 104859.
- Studer, R., Benjamins, V. R. & Fensel, D. (1998) Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1), pp. 161–197.
- Tamašauskaitė, G. & Groth, P. (2023) Defining a Knowledge Graph Development Process Through a Systematic Review. *ACM Transactions on Software Engineering and Methodology*, 32(1), p. 27:1-27:40.
- UK BIM Framework (2021) Information management according to BS EN ISO 19650 - Guidance Part 3: Operational phase of the asset life-cycle.
- Van Assche, D., Delva, T., Haesendonck, G., Heyvaert, P., De Meester, B. & Dimou, A. (2023) Declarative RDF graph generation from heterogeneous (semi-)structured data: A systematic literature review. *Journal of Web Semantics*, 75, January, p. 100753.
- Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J. & Wang, G. (2023) GPT-NER: Named Entity Recognition via Large Language Models. arXiv.
- Xie, X., Moretti, N., Merino, J., Chang, J. Y., Pauwels, P. & Parlikad, A. K. (2022) Enabling building digital twin: Ontology-based information management framework for multi-source data integration. *IOP Conference Series: Earth and Environmental Science*. IOP Publishing, 1101(9), p. 092010.
- Yahya, M., Breslin, J. G. & Ali, M. I. (2021) Semantic Web and Knowledge Graphs for Industry 4.0. *Applied Sciences*. Multidisciplinary Digital Publishing Institute, 11(11), p. 5110.
- Zhang, J., Chen, B., Zhang, L., Ke, X. & Ding, H. (2021) Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, 2, January, pp. 14–35.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y. & Wen, J.-R. (2023) A Survey of Large Language Models. arXiv.
- Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H. & Zhang, N. (2023) LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities, May.