Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

# Zaha: a RAG-based question answering system for the urban environment in The World Avatar

Xinhong Deng[1], Yi-Kai Tsai[1], Srishti Ganguly[1], Dan Tran[1], Hou Yee Quek[2], Wilson Ang[1], Shin Zert Phua[1], Sebastian Mosbach[1,2,3], Jethro Akroyd[1,2,3], Markus Kraft[1,2,3]

Draft of November 18, 2024

[1] CARES
Cambridge Centre for Advanced
Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

[2] Department of Chemical Engineering
and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

[3] CMCL
No. 9, Journey Campus
Castle Park
Cambridge
CB3 0AX
United Kingdom

Preprint No. 330

UNIVERSITY OF CAMBRIDGE

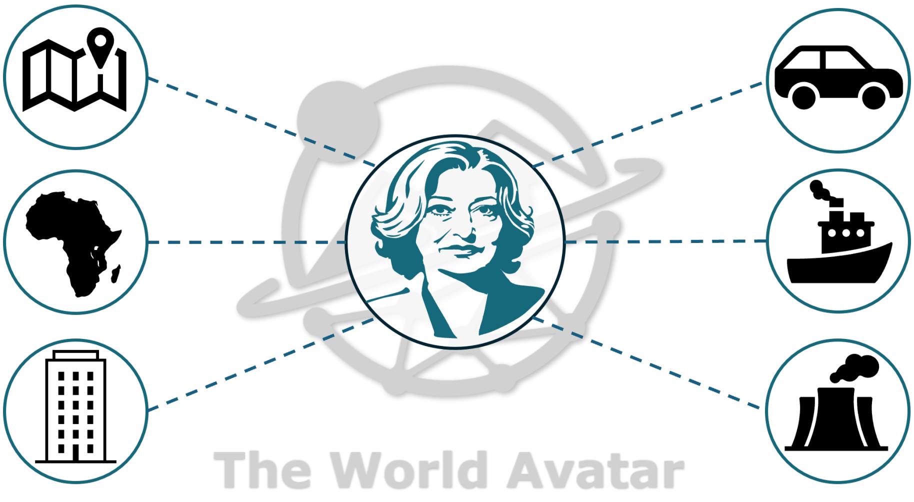
*Keywords:* urban planning, question answering systems, knowledge graph, large language models, retrieval-augmented generation

**Edited by**
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

**E-Mail:** mk306@cam.ac.uk
**World Wide Web:** https://como.ceb.cam.ac.uk/

## Abstract

As urbanisation accelerates, the demand for data-driven approaches to urban planning and management becomes increasingly critical. Urban data, encompassing geospatial, environmental, and regulatory information, is pivotal in informing the decision-making processes. However, various challenges such as data silos and the technical expertise required to query these complex datasets hinder the accessibility and utilisation of urban data. The World Avatar dynamic knowledge graph addresses these issues by integrating urban data across diverse sources and formats. Nonetheless, the process of querying from knowledge graphs remains non-trivial for non-experts due to the need for proficiency in query languages like SPARQL. This paper introduces 'Zaha', a retrieval augmented generation question answering system integrated with The World Avatar knowledge graph, enabling natural language queries for urban data. By leveraging knowledge graph technology, Zaha is capable of efficient data retrieval based on the relationships defined between entities. As such, Zaha enables users, including urban planners, to access and query complex urban data intuitively, bypassing the technical expertise barrier.



## Highlights

- A natural language question answering system for urban data
- Ontological knowledge representation of urban data
- Diverse output formats tailored for various urban applications

# Contents

# 1 Introduction

As urbanisation continues, the complexities of managing and planning modern cities have grown [48, 60]. Data-driven approaches for urban planning are increasingly gaining prominence [8, 9, 19, 38]. Urban data plays a critical role in informing the decision-making process, and encompasses a wide range of information such as geospatial, environmental, and regulatory datasets [48, 60]. The increasing reliance on urban data has also brought about several challenges. Data silos are one of the more significant barriers to effective city planning, as data often originates from various different sources and is stored in inconsistent formats [4, 6, 48, 60]. Additionally, urban data encompasses both static information, such as zoning regulations, and dynamic information, such as real-time air quality measurements, further complicating urban data management. The complexity of urban data structures, coupled with the need for technical expertise in query languages, often leaves valuable information inaccessible or underutilised, particularly by non-experts who may not be familiar with complex data analytics tools.

To address these interoperability issues, The World Avatar project introduces a dynamic knowledge graph approach to facilitate the seamless exchange and integration of diverse data and knowledge sources [2]. A knowledge graph is a structured representation of information, organised in a graph-like format that captures the relationships between entities [47]. By providing a standardised data model that defines entities and their relationships, knowledge graphs are able to integrate disparate data sources, thereby helping to resolve urban data silos. The dynamic aspect of The World Avatar refers to the knowledge graph evolving over time and continuously incorporating new data. Computation agents within The World Avatar serve as executable knowledge units that update the knowledge graph to maintain its dynamicity. These features are highly advantageous in urban planning, where up-to-date data from a wide range of domains must be synthesised to support decision-making.

While a dynamic knowledge graph enhances data integration, the process for non-experts to query for information from the knowledge graph could still be non-trivial as it would require proficiency in specific query languages such as SPARQL. To overcome this barrier, a knowledge graph-based question answering system can be employed, allowing users to retrieve information using natural language queries instead of forming complex formal queries. Knowledge graph-based question answering systems allow users to input and receive responses in natural language [27] bypassing the need for technical expertise, while leveraging the structured relationships of knowledge graphs for reasoning and inference [61]. This makes the data stored in knowledge graphs more accessible to a broader range of users, including urban planners, policymakers, and other stakeholders, who may lack the technical skills to work directly with the underlying knowledge graph structures. Furthermore, integrating a question answering system with The World Avatar dynamic knowledge graph ensures that the responses remain current and reflective of up-to-date information, thereby supporting dynamic and data-driven urban planning.

The **purpose of this paper** is to introduce an urban data question answering system integrated with The World Avatar dynamic knowledge graph, enabling intuitive access to the diverse urban data in the knowledge graph and supporting more effective urban planning and management. The proposed system addresses both data integration and access challenges often encountered when dealing with urban data, offering a streamlined, natural language approach for querying complex urban data without the need for technical expertise.

The structure of the paper is as follows: Section 2 presents an overview of urban data challenges, The World Avatar dynamic knowledge graph, and question answering systems. Section 3 outlines the development of urban data ontology relevant to this paper. Section 4 introduces an urban data question answering system within The World Avatar, designed to support urban planning and management. This section further details the technical implementation and an assessment of the robustness of the system. Section 5 discusses practical applications of the question answering system and Section 6 concludes the work.

# 2 Background

## 2.1 Urban data challenges in city planning

Urban data plays a critical role in city planning, providing insights for infrastructure, transportation, environment, and land use [7, 9, 60]. With increasing urbanisation, data-driven approaches are essential in streamlining information exchange between individuals and the city [7], and in enabling the city to make informed decisions on urban planning that could affect millions of people [6]. These approaches use different types of urban data, including geospatial, environmental, demographic, and regulatory, to manage the complexities of modern cities and improve the quality of life of residents [48, 60]. For instance, real-time air quality data helps shape public health policies in a city, particularly for industrial and maritime regions [48], while land use records and land suitability maps inform zoning and development regulations [60]. Additionally, building geometry data is crucial for evaluating energy optimisation and compliance with regulations [31]. These diverse datasets often need to be integrated to address specific problems, such as combining real-time parking availability data with city infrastructure data to help in parking space management, which is often a tedious task in densely populated urban environments [48, 50].

While datasets from various sectors are crucial for understanding urban dynamics, they are often stored in different systems and formats. Barbosa et al. [4] conducted a study on open datasets from cities in the United States and Canada, revealing that urban datasets are highly diverse and published in various formats, including tables, maps, charts, and documents. The data formats found include Comma-Separated Values (CSV), Resource Description Framework (RDF), Extensible Markup Language (XML), Portable Document Format (PDF), and Keyhole Markup Language Zipped (KMZ), amongst others. This issue is compounded by the nature of the data with some of it being static (*e.g.,* land use maps) and some dynamic (*e.g.,* real-time traffic or pollution data), containing both qualitative (*e.g.,* regulations) and quantitative (*e.g.,* population growth) elements [4, 6, 48, 51, 60]. Additionally, different terms are often used to describe the same attribute, and the same term can also represent multiple concepts at the same time [4]. The lack of standardisation in semantics [51] leads to difficulties in digitally integrating and processing information.

As a result, there is a need for scalable and automated methods for processing and integrating these datasets [4].

Even when the data is available, users often struggle to access and interpret it [58]. The underlying data models and software functions are often large and complex, and require significant time and effort to fully understand the data acquisition process [58]. For data on Geographic Information Systems (GIS), the reliance on different, incompatible software has caused users to adopt specific terminology and interface familiarity, making it difficult to transfer their knowledge to other systems for similar data [11]. Although multiple query languages are available for data extraction, their use requires users to have a deep understanding of the underlying data structures as well as the query languages [22]. It is also challenging to query regulatory data due to its textual nature, requiring users to read, search, and filter through information, which is both time-consuming and demanding of specialised knowledge [62]. This limits accessibility for multi-disciplinary users involved in city planning, such as civil engineers who may not fully grasp regulations intended for urban planners [62].

These challenges highlight the need for more accessible, streamlined methods for integrating, accessing, and understanding urban data, allowing for more efficient city management and decision-making.

## 2.2 The World Avatar dynamic knowledge graph

The World Avatar (TWA) project aims to create a digital representation of the world by utilising Semantic Web technologies and dynamic knowledge graphs. The objective of this approach is to facilitate the integration of diverse data sources and concepts, ensuring seamless interoperability across different domains [37]. Through the use of knowledge graphs, TWA addresses the challenges of urban data silos, enabling easy access to the complex data crucial for city planning. TWA has demonstrated its versatility through successful applications in various cross-domain problems [2], including knowledge graph-based question answering systems for the chemistry domain [32, 43, 52, 64].

At its core, TWA uses Resource Description Framework (RDF) and ontologies to structure its data into triples that represent relationships between entities in a `subject-predicate-object` format [33]. This structured data forms the basis of directed knowledge graphs, where all of the components are represented with unique Internationalised Resource Identifiers (IRIs) [17] to ensure an interconnected data structure and facilitate easy retrieval, identification, and linking of information across diverse datasets. An ontology is a formal, explicit specification of a particular domain of knowledge, defining the entities, relationships, and rules within that domain [21]. In this paper, we discuss various ontologies related to cities, defining key entities such as buildings, land plots, regulations, and environmental conditions, along with their relationships. Ontologies make the data semantically rich, ensuring they can be consistently queried due to their standardised structure [55].

TWA knowledge graph is also dynamic in nature, indicating that it continuously evolves through the integration of new data. This is achieved using various autonomous software components called ‘agents’, which are fully integrated into TWA with semantic description and regularly update the knowledge graph with information from diverse sources [25, 63]. These agents can monitor real-time data, such as parking lot availability, and update the current time series in the knowledge graph. Additionally, the agents can also conduct complex tasks, including running simulations using various external tools, to provide insights into urban interactions, such as simulating the pollution levels generated by ships around Singapore. Hence, TWA's network of agents is crucial in maintaining the accuracy and relevance of the data in the knowledge graph.

To make the knowledge graph data accessible, TWA employs triplestores, such as Blazegraph [10] and RDF4J [18], to store data directly in the form of triples. In addition, TWA utilises virtual knowledge graphs like Ontop [59], which store data in relational databases in tabular formats, allowing users to define mappings between the relational data and RDF triples without the need to physically transform the data. The data is then made available for querying using the standard SPARQL Protocol and RDF Query Language [57], enabling users to execute complex queries to integrate heterogeneous datasets effectively.

## 2.3 Question answering systems

Question answering (QA) systems are powerful platforms designed for automatically answering questions asked in natural language [12]. Interest in QA systems has grown due to their capability to provide question-specific answers in natural language [56]. QA systems can vary in their approach, and knowledge graph-based QA (KGQA) systems have seen an increase in prominence due to their ability to leverage structured, semantic representation of data [3, 61]. Knowledge graphs, combined with ontologies, allow for the integration of information across different domains and sources while defining relationships between entities. This allows for reasoning over the data and enhances the ability of QA systems to handle queries more effectively.

A key component of QA systems is natural language processing (NLP), which concerns the use of computational techniques to learn, understand and generate human language content [23]. One of the most prominent approaches to NLP in recent years has been the use of different types of neural networks, specifically large language models (LLMs) [39]. LLMs are effective in understanding complex natural language inputs and generating fluid, human-like responses. In the context of KGQA systems, LLMs can be used to translate natural language questions into SPARQL queries [20, 44]. This capability frees users from having to know SPARQL syntax, allowing them to access the knowledge graph data with natural language inputs. By lowering the technical barrier, LLM-enhanced KGQA systems make querying more accessible, while still delivering relevant answers based on the structured data stored in knowledge graphs.

In-context learning (ICL) plays an important role in enhancing the performance of LLMs within QA systems. ICL is a paradigm for NLP, where LLMs generate predictions based on context, which are augmented with examples, without needing fine-tuning for each task [16]. This greatly reduces the computational costs and allows LLMs to adapt dynamically to new tasks. Retrieval-augmented generation (RAG) is a method of ICL that enables LLMs to have access to new knowledge sources [46]. RAG helps reduce hallucination by supplying LLMs with relevant information [26], allowing them to adapt dynamically to new information, making the model more versatile across different domains and up-to-date with evolving information. RAG comprises two main components, retrieval and generation [35]. The retrieval component is responsible for gathering relevant facts from external data sources based on the input question. The generator component, powered by LLM, takes the retrieved information to generate contextually accurate natural language responses to the question.

There have been multiple QA systems and LLM applications developed in the various domains related to urban planning [15, 36, 40, 58, 62]. These systems address a wide range of tasks, from geospatial analysis, regulatory compliance to building information modelling (BIM). Existing solutions tend to address only specific domains or tasks. Li and Ning [36] developed an autonomous GIS leveraging LLM to address various geospatial tasks. This system requires users to provide input data files and relevant metadata and produces responses for questions related to the input files. Zhong et al. [62] developed a QA system specifically designed for the domain of building regulations, where the system uses a deep learning model of NLP to retrieve information from a collection of building regulation documents, and produces natural language responses to questions. Wang et al. [58] developed a QA system for BIM to support decision-making by construction project team members. An Industry Foundation Classes (IFC) file is inputted to the system, and the system responds in natural language to questions related to the file asked by the user. DC Compass [15] is a tool leveraging NLP to understand and respond to user questions regarding public data from the District of Columbia's Open Data platform [1]. While the tool is able to retrieve public data, it does not perform any computation or reasoning on the data to answer questions.

Despite the growth in QA systems within urban planning, existing solutions remain focused on specialised domains or functions, and may only support certain input data formats or output formats. There is a need for a comprehensive QA system that integrates the diverse urban planning domains, handles the various types of input data, and has the capability of generating multiple output forms. Such a system would streamline the access to heterogeneous information sources and enhance decision-making in urban planning.

# 3 Urban data ontology developments

Ontologies play a central role in the work presented in this paper, as they allow for intelligent data integration and retrieval across various domains. **Figure** 1 outlines the key concepts and properties of the various ontologies used in this study.

OntoDispersion is an existing ontology designed for dispersion simulation data [25]. It includes a geospatial concept, `od:Scope`, which is defined as a subclass of `geo:Feature`, enabling geospatial capabilities via GeoSPARQL. The dispersion data are stored as raster in a PostGIS database, with the metadata captured by `od:DispersionRaster` instances linked to `od:DispersionOutput` instances. This structure allows agents to efficiently retrieve dispersion data by first querying for the metadata via SPARQL, followed by an SQL query to obtain the underlying raster value from the PostGIS database.

OntoBuiltEnv is an existing ontology developed to represent key information and properties of the built environment [24]. It incorporates the `obe:PropertyUsage` concept

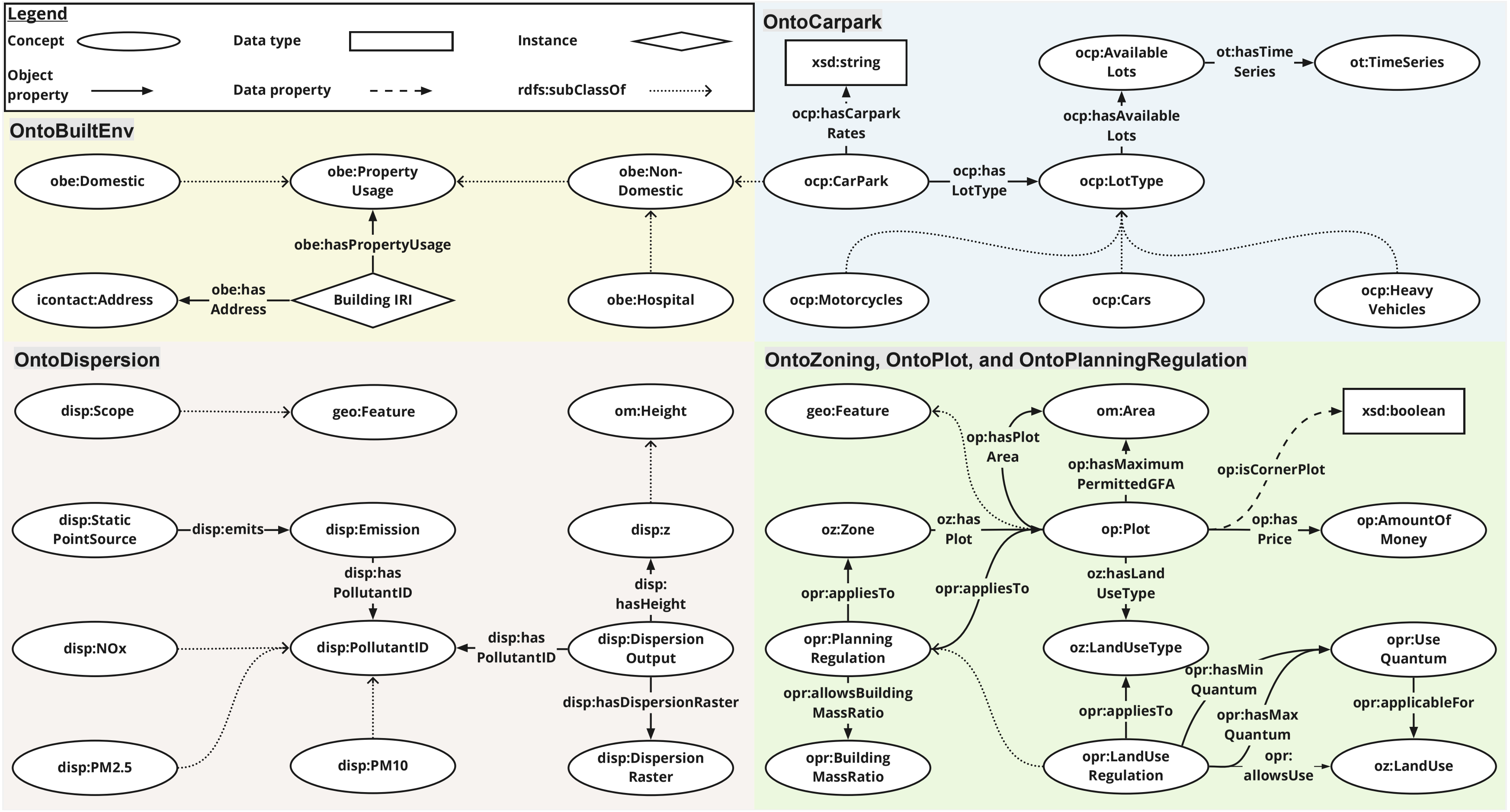


**Figure 1:** *Overview of ontologies used in this work in the domain of urban data: OntoDispersion, OntoBuiltEnv, OntoPlot, OntoPlanningRegulation, OntoZoning, OntoCarpark.*

for the representation of building usage, and captures essential construction and market value details derived from various publicly available data sources. To ensure interoperability, OntoBuiltEnv also reuses existing concepts from other ontologies, such as `icontact:Address` from iCity address ontology [29], for detailed address information.

Beyond leveraging existing ontologies, we also extended existing ontologies and developed new ones to support the work done in this paper. OntoZoning is an existing ontology developed for representing the relationships between zones, land uses, and regulations [49]. Although it successfully captured the necessary concepts, its size and complexity led to challenges in management and efficient use. To mitigate these issues, in this paper, we split the ontology into three smaller, more focused domain ontologies: OntoPlot, OntoPlanningRegulation and a streamlined version of OntoZoning. This separation enhances modularity, allowing users to focus on the ontology that is most relevant to their use cases, while improving maintainability.

OntoPlot is developed to describe the attributes of land plots. It includes a central `op:Plot` concept, which is defined as a subclass of `geo:Feature`, enabling GeoSPARQL queries. Key attributes providing detailed description of the plots are defined as object and data properties, including `op:hasMaximumPermittedGFA`, `op:hasPlotArea` and `op:isCornerPlot`.

OntoPlanningRegulation handles urban planning regulations. There are two types of regulations, both of which are included in the ontology: the standard planning regulation `opr:PlanningRegulation`, and the land use regulation `opr:LandUseRegulation`, which is a subclass of `opr:PlanningRegulation`. A `opr:PlanningRegulation` can have different rules, such as `opr:allowsBuildingMassRatio`, while `opr:LandUseRegulation` introduces further constraints, such as `opr:allowsMinUseQuantum` and `opr:allowsMaxUseQuantum`.

The revised OntoZoning presented in this paper focuses on the `oz:Zone` concept, also a subclass of `geo:Feature`, allowing for GeoSPARQL capabilities. `oz:Zone` is associated with `op:Plot` via `oz:hasPlot`. OntoZoning further includes the `oz:LandUseType` concept, where each `op:Plot` `oz:hasLandUseType` `oz:LandUseType`. `op:LandUseRegulation` `opr:appliesTo` each `oz:LandUseType`. The regulations further specify which `oz:LandUse` is applicable for the different `oz:LandUseType` via `opr:allowsUse`. For example, `oz:Park` `opr:allowsUse` `oz:CommunityPark` and `oz:Park` `opr:allowsUse` `oz:NationalPark`. Various sub-classes of `oz:LandUseType` and `oz:LandUse` are also defined in the ontology, covering a wide range of use cases.

In this work, an ontology for car parks is also required to represent essential urban information. The iCity parking ontology [30] provides a framework for representing parking areas and includes key information, such as the parking rate `ipark:ParkingRate`. However, it lacks concepts to capture certain details about the car parks, such as the number of available parking lots in real time. Furthermore, the iCity parking ontology does not integrate with OntoBuiltEnv, which would have allowed for GeoSPARQL capabilities. To address these gaps, we developed the OntoCarpark ontology to providea a more comprehensive representation of car park details. The central concept `ocp:Carpark` is defined as a subclass of `obe:Non-Domestic`, which is a subclass of `obe:PropertyUsage`, to ensure compatibility and integration with OntoBuiltEnv. OntoCarpark allows querying key car park attributes, such as real-time availability using the object property `ocp:hasAvailableLots` and rates via `ocp:hasCarparkRates`. It also accommodates more granular details, differentiating rates across different days using data properties like `ocp:hasWeekdayRates`, `ocp:hasSaturdayRates`, and `ocp:hasSundayAndPHRates`.

The development of these ontologies enhances the representation and integration of urban data in TWA dynamic knowledge graph, ensuring that detailed and structured information is available for analysis and decision-making.

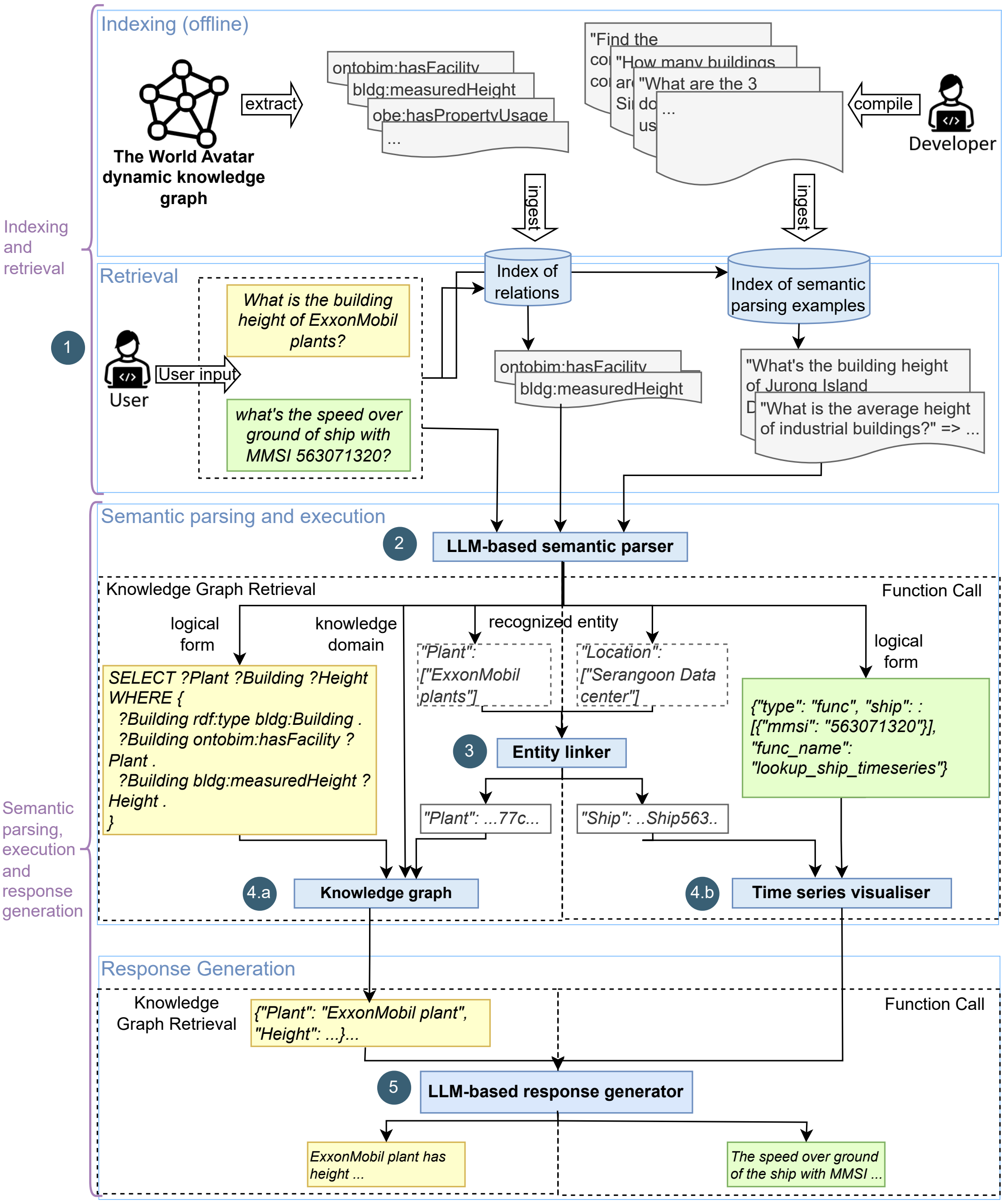
# 4 Question answering system for urban data

To support querying of urban data, we have developed Zaha, a QA system that integrates various ontologies and datasets from TWA's dynamic knowledge graph. Zaha builds on the architecture of our existing chemistry QA system, Marie [52], providing a unified, natural language access point to a range of urban data domains through a RAG-based architecture. While Marie effectively handled chemistry-related queries, it was limited in addressing the diverse data formats and complex output requirements found in urban data. Zaha overcomes these challenges by leveraging TWA's software agents, enabling it to perform advanced functionalities such as generating time series plots and visualising building footprints, thereby expanding the QA system's application in urban data contexts.

## 4.1 System implementation

Zaha consists of three main modules: indexing and information retrieval, semantic parsing and execution, and chat generation. It uses *all-mpnet-base-v2* [28] for text embedding, OpenAI's *gpt-4o-mini* model [42] for in-context learning and Redis Community Edition [45] for vector search and retrieval. As shown in **Fig.** 2, Zaha's architecture is structured around the vectorisation and indexing of data from both the knowledge graph and a curated dataset to enable faster information retrieval. When a user submits a question, relevant data is retrieved and fed into the LLM along with the input question for semantic parsing, to generate a logical form.

To address dynamic questions and offer diverse output formats, the semantic parsing generates two types of logical forms: SPARQL query for accessing the knowledge graph, and metadata retrieval or function call to external tools. **Figure** 3 illustrates Zaha's workflow and highlights the differences between these two types of logical forms, demonstrated with two example questions: "What is the percentage of plots zoned for residential use, out of the total number of plots?" and "What is the speed over ground of ship with MMSI 563071320?" The action execution results, returned in a JSON response, are then visualised and used by the LLM to generate a natural language answer.

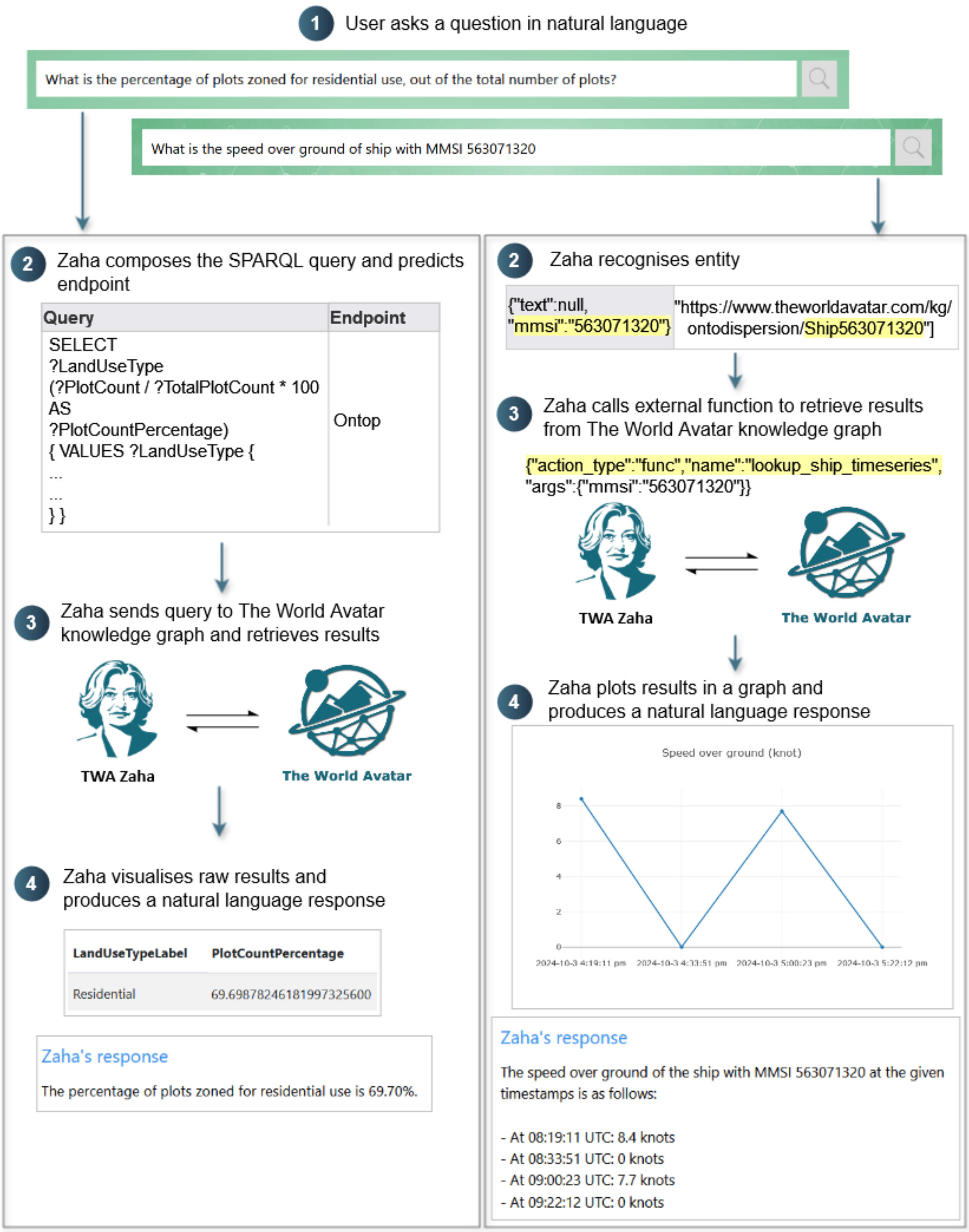

Indexing and retrieval
Indexing (offline)
The World Avatar dynamic knowledge graph
extract
ontobim:hasFacility
bldg:measuredHeight
obe:hasPropertyUsage
...
"Find the
"How many buildings
"What are the 3
...
compile
Developer
ingest
ingest
1
Retrieval
User
User input
What is the building height of ExxonMobil plants?
what's the speed over ground of ship with MMSI 563071320?
Index of relations
Index of semantic parsing examples
ontobim:hasFacility
bldg:measuredHeight
"What's the building height of Jurong Island
"What is the average height of industrial buildings?" => ...
Semantic parsing, execution and response generation
Semantic parsing and execution
2
LLM-based semantic parser
Knowledge Graph Retrieval
Function Call
logical form
knowledge domain
recognized entity
logical form
"Plant": ["ExxonMobil plants"]
"Location": ["Serangoon Data center"]
3
Entity linker
"Plant": ...77c...
"Ship": ..Ship563..
4.a
Knowledge graph
4.b
Time series visualiser
Response Generation
Knowledge Graph Retrieval
Function Call
5
LLM-based response generator
ExxonMobil plant has height ...
The speed over ground of the ship with MMSI ...

```
SELECT ?Plant ?Building ?Height
WHERE {
  ?Building rdf:type bldg:Building .
  ?Building ontobim:hasFacility ?
Plant .
  ?Building bldg:measuredHeight ?
Height .
}
```

```
{"type": "func", "ship": :
[{"mmsi": "563071320"}],
"func_name":
"lookup_ship_timeseries"}
```

```
{"Plant": "ExxonMobil plant",
"Height": ...}...
```

**Figure 2:** *Overview of Zaha's architecture, consisting of one offline indexing stage and three online stages: retrieval, semantic parsing and execution, and response generation. The workflow proceeds as follows: (1) The user asks Zaha a question. (2) The LLM generates logical forms, identifying relevant entities, and necessary details. (3) Zaha performs entity linking to match entities in the question to their IRIs. (4) Based on the logical forms, Zaha either (a) executes a SPARQL query or (b) calls specific functions. (5) The LLM then generates a natural language response to the question, using the results from the previous step.*



1 User asks a question in natural language

What is the percentage of plots zoned for residential use, out of the total number of plots?

What is the speed over ground of ship with MMSI 563071320

2 Zaha composes the SPARQL query and predicts endpoint

| Query | Endpoint |
|---|---|
| SELECT<br>?LandUseType<br>(?PlotCount / ?TotalPlotCount * 100<br>AS<br>?PlotCountPercentage)<br>{ VALUES ?LandUseType {<br>...<br>...<br>} } | Ontop |

3 Zaha sends query to The World Avatar knowledge graph and retrieves results

4 Zaha visualises raw results and produces a natural language response

| LandUseTypeLabel | PlotCountPercentage |
|---|---|
| Residential | 69.69878246181997325600 |

Zaha's response

The percentage of plots zoned for residential use is 69.70%.

2 Zaha recognises entity

```
{"text":null,
"mmsi":"563071320"}
```

```
"https://www.theworldavatar.com/kg/
ontodispersion/Ship563071320"]
```

3 Zaha calls external function to retrieve results from The World Avatar knowledge graph

```
{"action_type":"func","name":"lookup_ship_timeseries",
"args":{"mmsi":"563071320"}}
```

4 Zaha plots results in a graph and produces a natural language response

Zaha's response

The speed over ground of the ship with MMSI 563071320 at the given timestamps is as follows:

- At 08:19:11 UTC: 8.4 knots
- At 08:33:51 UTC: 0 knots
- At 09:00:23 UTC: 7.7 knots
- At 09:22:12 UTC: 0 knots

**Figure 3:** *Overview of Zaha's workflow, highlighting the generation of SPARQL queries and external function calls in response to diverse questions, exemplified by queries related to land use and ship tracking.*

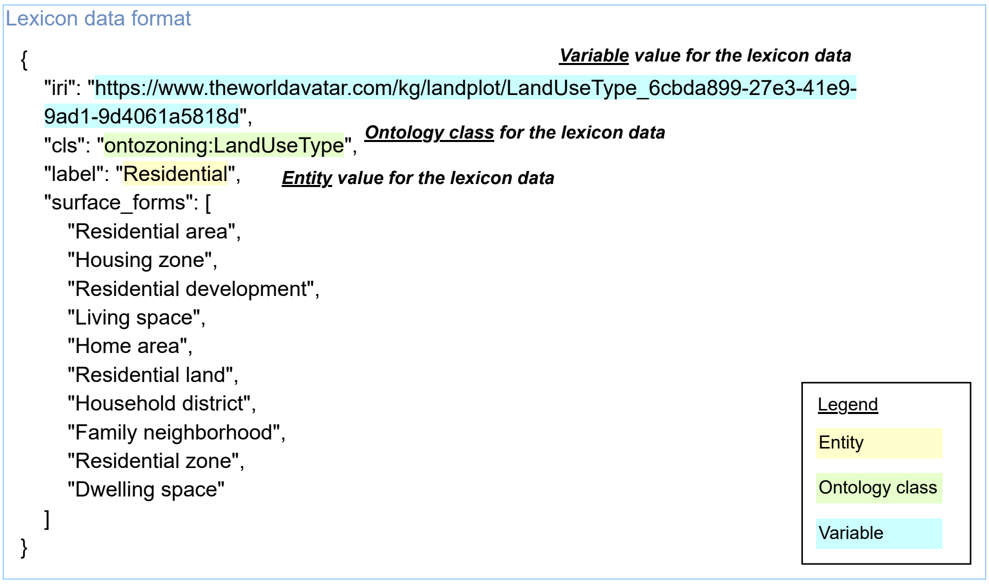
### 4.1.1 Indexing and retrieval

Indexing and retrieval are foundational steps in a RAG system, providing the context needed by the LLM for answer generation and directly influencing the overall effectiveness of the QA system. To streamline the retrieval of complex information, Zaha combines reasoning capabilities and semantic technologies with data from TWA knowledge graph and curated domain-specific datasets. These components are dynamically updated based on the user's question. Urban data-related ontology concepts are extracted from the TWA knowledge graph and indexed for efficient retrieval, while the curated datasets are also indexed to further enhance the process. These indexed datasets contain mappings between natural language questions, corresponding ontology concepts, and the associated SPARQL queries or function calls. **Figure** 4 presents two data formats used in indexing process, where **Fig.** 4(a) shows an example of knowledge extracted from knowledge graph and **Fig.** 4(b) is a data instance from the curated domain-specific dataset.

As shown in **Fig.** 4(a), the lexicon information of various entities is exported from the knowledge graph as JavaScript Object Notation (JSON) objects. The *label* represents the standard name of an entity, identified by its *iri* in the knowledge graph, and the entity can have multiple *surface_forms*, which are alternative names of the entity in text. The *cls* denotes the base class of the entity, which is used to narrow the search range during entity linking, considerably improving response time.

The curated datasets contain standard example questions in urban data to guide the LLM in generating complex queries, such as calculation, aggregation, filtering, and multi-levels graph traversal, offering a more efficient approach than random walk exploration. **Figure** 4(b) illustrates an example of the sample question structure, where a natural language question (*nlq*) is mapped to a logical form (*data_req*). Two key components, *entity_bindings* and *var2cls*, are employed during entity linking. *entity_bindings* identifies the entities in the natural language question and their corresponding variables in SPARQL query, while *var2cls* maps the entity variables to the ontology class. These ontology classes are aligned with the *cls* value in the lexicon data of Fig. 4(a) to facilitate faster entity searches. *req_form* defines the type of action and its content needed to answer the natural language question. A more detailed explanation of entity linking can be found in Section 4.1.2.

During the retrieval stage, Zaha uses indexed information to align the input question with pertinent ontology concepts, SPARQL query structures, or function call patterns. Zaha performs a vector similarity search to retrieve example questions by embedding the input question, as a vector and comparing it to vectors of pre-indexed sample questions. The ten most similar examples are selected based on similarity rankings and passed to the LLM, along with the input question for semantic parsing and answer generation. These sample questions are ranked and added to the prompt in descending order, ensuring that the LLM first attends to the most relevant examples while also drawing from a broader range of less similar examples to support a diverse contextual understanding.



Lexicon data format

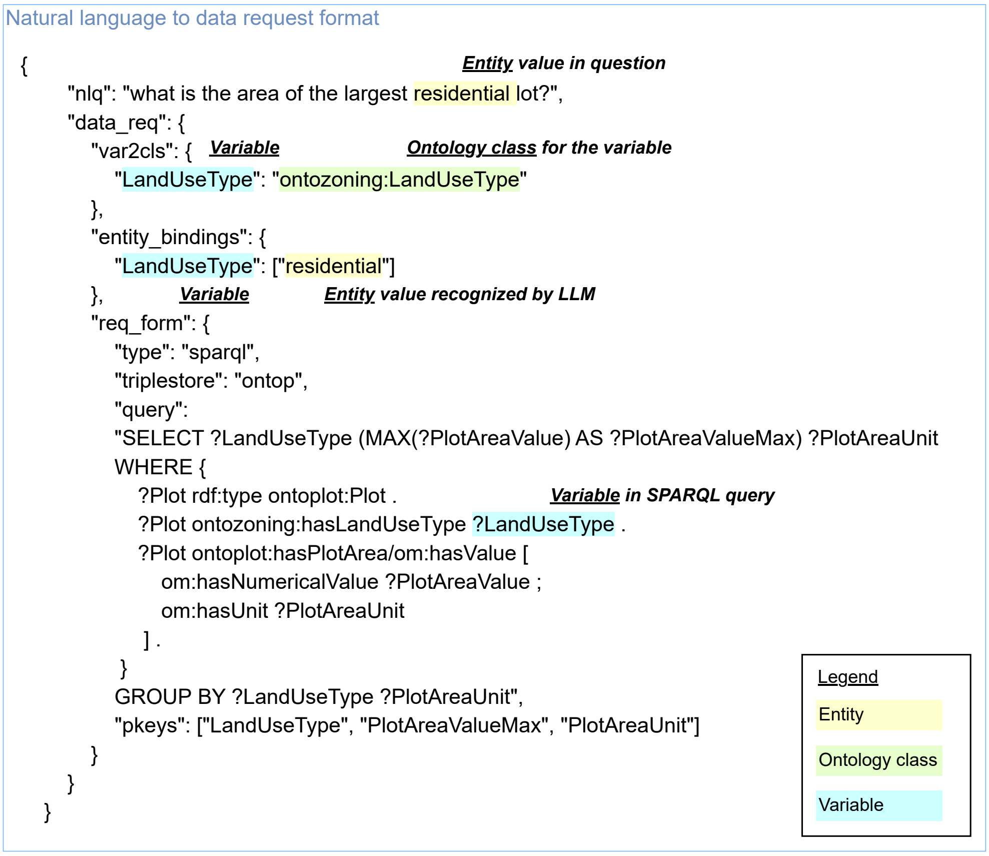
```
{
   "iri": "https://www.theworldavatar.com/kg/landplot/LandUseType_6cbda899-27e3-41e9-
   9ad1-9d4061a5818d",
   "cls": "ontozoning:LandUseType",
   "label": "Residential",
   "surface_forms": [
      "Residential area",
      "Housing zone",
      "Residential development",
      "Living space",
      "Home area",
      "Residential land",
      "Household district",
      "Family neighborhood",
      "Residential zone",
      "Dwelling space"
   ]
}
```

***Variable* value for the lexicon data**

***Ontology class* for the lexicon data**

***Entity* value for the lexicon data**

(a) An example of lexicon data for the entity *Residential*, containing its IRI, ontology class, label and alternative names.



Natural language to data request format

```
{
      "nlq": "what is the area of the largest residential lot?",
      "data_req": {
         "var2cls": {
            "LandUseType": "ontozoning:LandUseType"
         },
         "entity_bindings": {
            "LandUseType": ["residential"]
         },
         "req_form": {
            "type": "sparql",
            "triplestore": "ontop",
            "query":
            "SELECT ?LandUseType (MAX(?PlotAreaValue) AS ?PlotAreaValueMax) ?PlotAreaUnit
            WHERE {
               ?Plot rdf:type ontoplot:Plot .
               ?Plot ontozoning:hasLandUseType ?LandUseType .
               ?Plot ontoplot:hasPlotArea/om:hasValue [
                  om:hasNumericalValue ?PlotAreaValue ;
                  om:hasUnit ?PlotAreaUnit
                ] .
             }
            GROUP BY ?LandUseType ?PlotAreaUnit",
            "pkeys": ["LandUseType", "PlotAreaValueMax", "PlotAreaUnit"]
         }
      }
   }
```

***Entity* value in question**

***Variable***

***Ontology class* for the variable**

***Variable***

***Entity* value recognized by LLM**

***Variable* in SPARQL query**

(b) An example of converting a natural language question into a data request format, showing the question and data request details, including entity linking information (*var2cls*, *entity_bindings*, and *req_form*), to answer "What is the area of the largest residential lot?"

**Figure 4:** *Overview of indexed data format.*

### 4.1.2 Semantic parsing, execution and generation

Semantic parsing transforms the natural language question into logical forms by using both the user input query and the relevant information retrieved during the indexing and retrieval step. The LLM generates a data request in the format shown in Fig. 4(b). Zaha currently supports two types of request forms: SPARQL queries for accessing the TWA knowledge graph and function calls to external tools, such as TWA software agents, to perform computations or retrieve additional data.

Entity linking is then performed to match the entities in the user input question with their corresponding IRIs, which are used for information retrieval or function calls. The LLM identifies the entity, variables for the entity used throughout the search process, and the actual ontology class from the input. These classes and values are subsequently used for IRI matching. For example, in Fig. 4(b), *residential* is identified as a *LandUseType* entity class, which has the ontology class *ontozoning:LandUseType*. During IRI matching, the query engine needs to search for a match in the lexicon data, where *cls* field has value *ontozoning:LandUseType* and *label* field has value *residential*. The IRI identified from this process is then used for subsequent operations.

Zaha offers multiple IRI matching strategies for individual classes, including *direct match, semantic search, fuzzy search, and customised linking*. In *direct match*, the system identifies the IRI by searching the lexicon for exact literal matches to the surface form. *Semantic search* [5], a method for finding terms that are semantically similar, uses an LLM-based embedder to transform the entity text into a vector, capturing its semantic meaning, and then applies k-nearest neighbour (KNN) [14] to find the closest matches. *Fuzzy search* [41] addresses typographical errors in natural language input by also embedding the entity text with the LLM, but instead of KNN, it uses Levenshtein distance [34] to account for ambiguity. This search method is commonly applied to named entities, such as building names, which typically carry minimal semantic meaning and are prone to typographical errors. The *customised linker* handles cases where class-specific processing is required for certain entity classes. During the entity linking process, multiple mappings can be identified and passed to the execution unit for further analysis.

For each question, Zaha starts by determining if a customised linker is needed. If so, it runs the customised linker and then exits the entity-linking process. Otherwise, Zaha first attempts a direct match for entity linking. If a direct match does not yield sufficient results, it proceeds with either semantic search or fuzzy search, depending on the question type. The choice between semantic and fuzzy search is predefined based on the nature of the question, ensuring that the most contextually suitable matching strategy is applied for accurate linking.

Like most QA systems, information retrieval and reasoning are the core functionalities in Zaha. Once the recognised entities are linked to their corresponding IRIs, these IRIs are used to populate the SPARQL query, replacing the variables for the relevant entity classes. For instance, in Fig. 4(b), the variable *?LandUseType* is replaced with the IRI for *residential*. The SPARQL query is then executed by sending it to the appropriate knowledge graph endpoint, as predicted by Zaha.

Beyond text retrieval, Zaha can handle dynamic data and support various output formats

**Table 1:** *Current supported functions of Zaha.*

| Function | Input | Description |
|---|---|---|
| get_pollutant_conc | Location | Get pollution of the provided location |
| lookup_ship_timeseries | Ship IRIs | Get the timeseries data (*e.g.* speed over ground, course over ground, longitude and latitude) for the provided ship |
| find_nearest_carpark | Location | Get the nearest car park of the provided location and its availability |

in urban domains through external function calls. For urban-specific use cases, Zaha generates and executes a set of predefined functions rather than relying solely on SPARQL queries. These functions are invoked using the matched entity IRIs. Table 1 lists the current functions supported by Zaha.

Finally, the user input question, semantically parsed query, and the query execution results are passed to the LLM-based response generator to produce an answer in natural language.

## 4.2 System assessment

We evaluated the effectiveness of Zaha's architecture by focusing on the accuracy of operation generation rather than the factual accuracy of the answers themselves. This assessment verifies Zaha's capability to reliably retrieve relevant information in response to user questions. Since the accuracy of answers ultimately depends on the data stored in the knowledge graph, which can be dynamically updated, this approach ensures that Zaha's performance remains consistent as data evolves.

The retrieval process in Zaha is evaluated across three core components: entity linking, SPARQL generation, and function call accuracy. For entity linking, we assess Zaha's ability to correctly identify entities in user questions and match them with the appropriate IRIs in the knowledge graph. In SPARQL generation, we verify that Zaha constructs accurate SPARQL queries to retrieve contextually relevant data. For function calls, we examine whether Zaha can identify and execute the appropriate function with correct parameter values to address the user's query. Evaluating these components validates the robustness of Zaha's retrieval mechanism, which is essential to its overall performance.

### 4.2.1 Entity linking

When a user submits a question, Zaha begins by identifying possible entities within the text and linking them to the appropriate IRIs in the knowledge graph. We evaluate Zaha's linking performance across diverse scenarios, including cases where the question contains a fully accurate entity name, a semantically similar name, an entity name with typographical errors, or multiple entities. This approach ensures Zaha's linking process remains reliable and adaptable, accommodating a wide range of query types. Examples for each scenario can be found in Table 2. The assessment focuses on two key aspects: correctly

**Table 2:** *Examples for entity linking test scenarios.*

| Scenario | Input question | Expected entity | Expected IRI metadata |
|---|---|---|---|
| Complete entity name | What does Special Use mean as a land use category? | Special Use | `rdfs:label "Special Use"` |
| Semantically similar entity name | What land use category would hospitals be classified as? | hospitals | `rdfs:label "Health and Medical Care"` |
| Entity entity name with typographical errors | What's the building height of APEC INDSTRIES? | APEC INDSTRIES | `rdfs:label "APEC INDUSTRIES"` |
| Multiple entities | What's the difference between Business 1 and Business 2 land use classifications? | Business 1, Business 2 | `rdfs:label "Business 1", rdfs:label "Business 2"` |

identifying the entity in the question and linking it to the appropriate IRI. For the latter, accuracy is determined by verifying if the IRI's metadata, such as the `rdfs:label` or other attached properties, corresponds precisely to the intended entity.

For instance, in the question, "What is the building height of APEC Industries?" Zaha should identify "APEC Industries" as the entity and link it to an IRI with `rdfs:label "APEC INDUSTRIES (S) PTE. LTD."`. An incorrect identification would mean failing to find the entity, while incorrect IRI linking would occur if the system linked to an IRI with an incorrect label like `rdfs:label "MITSUNOBU INDUSTRIES PTE. LTD."` instead. This evaluation ensures the entity linking aligns with user intent and reinforces Zaha's ability to provide accurate responses.

The results of the entity linking assessment are summarised in Table 3, grouped by the different scenarios. As shown in the table, Zaha is highly accurate in identifying the entity in the question and correctly linking the entity to its respective IRI. The test questions used in this assessment can be found in Appendix A.2.

### 4.2.2 SPARQL generation

This section assesses the accuracy of the SPARQL query generation process in Zaha. Once the entity is identified and linked to an IRI, Zaha generates a SPARQL query to retrieve the relevant data. The process involves two main steps: first, Zaha creates a SPARQL template without the IRI; then, the IRI is inserted into this template to produce the final SPARQL query. The focus here is on assessing the accuracy of the SPARQL template generation, as the IRI insertion is a straightforward string replacement.

The assessment is divided into two parts, syntax verification and intended result retrieval rate. Syntax verification refers to whether the generated SPARQL has the correct syntax,

**Table 3:** *Entity linking assessment results summary.*

| Scenario | Number of questions | Entity identification (%) | IRI correctness (%) |
|---|---|---|---|
| Complete entity name | 22 | 99.99 | 99.99 |
| Semantically similar entity name | 19 | 99.99 | 99.99 |
| Entity entity name with typographical errors | 5 | 99.99 | 99.99 |
| Multiple entities | 4 | 99.99 | 99.99 |
| **Weighted Average** | | 99.99 | 99.99 |

which can be verified by checking if the SPARQL can be executed by a triplestore without causing errors. The second part assesses whether the SPARQL query is generated with the appropriate context to the question, ensuring that the intended result can be accurately retrieved.

For example, if a user asks "What is the building height of Singapore data centre?", the correct SPARQL generated would be

```
PREFIX rdf: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bldg: <http://www.opengis.net/citygml/building/2.0/>
PREFIX ontobim: <https://www.theworldavatar.com/kg/ontobim/>

SELECT ?DataCentre ?Building ?Height WHERE {
  ?Building rdf:type bldg:Building .
  ?Building ontobim:hasFacility <https://example.org/datacentre_iri> .
  ?Building bldg:measuredHeight ?Height .
}
```

where `<https://example.org/datacentre_iri>` is the IRI of Singapore data centre identified during the entity linking stage. The example SPARQL query generated has the correct syntax and can be successfully executed by a triplestore. The SPARQL also contains the relevant predicate, `bldg:measuredHeight`, which will retrieve the height of the building.

In the assessment of SPARQL generation, two question categories were tested: land use and urban data. The results, summarised in Table 4, indicate an accuracy of over 93.85% for syntax correctness and over 92.86% for the intended result retrieval rate. Syntax errors, when they do occur, typically involve more complex queries where issues may arise in the placement or the correctness of the `GROUP BY` clauses. For the retrieval rate, occasional issues arose from Zaha's incomplete understanding of the relevant ontologies, such as failing to recognise essential subclasses within the ontology needed to retrieve the intended data.

**Table 4:** *Topic-based SPARQL generation assessment results summary.*

| Topic | Number of questions | Syntax correctness (%) | Intended result retrieval rate (%) |
|---|---|---|---|
| Land use in Singapore | 40 | 90.00 | 90.00 |
| Urban data | 30 | 99.99 | 96.67 |
| **Weighted average** | | 93.85 | 92.86 |

### 4.2.3 Function call

This section evaluates Zaha's capability to perform dynamic data operations through function calls, in addition to static data retrieval. It examines two key aspects of robustness: first, whether Zaha correctly identifies and calls the appropriate function that provides a relevant solution to the user's query, and second, whether Zaha accurately provides the necessary input parameters to the called function.

As an example, consider the question, "What's the speed over ground of ship with MMSI 563071320?". In this case, the correct function for Zaha to call would be *lookup_ship_timeseries*, which retrieves the time series for the specified ship IRI. The appropriate input for this function would be the IRI corresponding to the ship with MMSI 563071320. If Zaha calls a different function or provides an incorrect IRI that does not correspond to the ship with MMSI 563071320, the function call is evaluated as incorrect.

Sixteen different questions were assessed, and the results of the function call assessment are shown in Table 5. The questions are grouped by topic, based on the functions currently supported by Zaha. As shown in Table 5, Zaha has a weighted average of 99.99% accuracy in calling the correct functions and supplying the appropriate input parameters.

# 5 Application areas

## 5.1 Land use in Singapore

Zaha enables natural language querying of qualitative, quantitative, and regulatory data to support land use planning, effectively addressing critical challenges in urban planning.

**Table 5:** *Function call assessment results summary.*

| Topic | Number of questions | Correct function call (%) | Correct input supplied (%) |
|---|---|---|---|
| Air pollution | 5 | 99.99 | 99.99 |
| Ship information | 5 | 99.99 | 99.99 |
| Carpark information | 6 | 99.99 | 99.99 |
| **Weighted average** | | 99.99 | 99.99 |

As illustrated in **Fig.** 5, users can pose qualitative questions to Zaha about land use in Singapore, such as “What land use categories would funeral parlours or columbaria be under?”. These queries yield natural language answers that are easily comprehensible. Users can also ask quantitative questions that require calculations or data summarisation, such as “What is the average Gross Floor Area allowed for residential plots in Singapore?” or “What is the area of the largest agricultural lot?”.

This approach also supports the implementation of regulation compliance checks. The interconnected structure of the knowledge graph allows for the integration of diverse datasets, such as building geometries and land use regulations. This enables users to ask regulatory questions like, “How many plots exceed the maximum permitted Gross Floor Area (GFA) for each land use type?”. To answer this question in a semantic manner, SPARQL queries can be constructed to extract the relevant information in the necessary order. The GFA of buildings on a given land plot is estimated using the building geometry data sourced from the knowledge graph. By summing the estimated GFA of all buildings on a plot and comparing this total against the applicable GFA regulations for that plot, it is possible to assess compliance with land use requirements.

The results of this query are presented in Fig. 5. While the data indicate numerous violations of land use regulations, it is important to recognise that these findings may not fully represent the ground truth; rather, they serve as a proof of concept to demonstrate Zaha’s capability to extend beyond basic queries by connecting various data domains and assessing regulatory compliance. This is because the GFA calculation is simplistic, as it solely considers the building’s external geometry and does not account for other structural elements that may influence GFA, such as balconies, overhangs, or mechanical penthouses [54]. Furthermore, the regulations applied do not encompass all exceptions to the rules; for instance, buildings featuring sky gardens are typically permitted a higher GFA than usual [53].

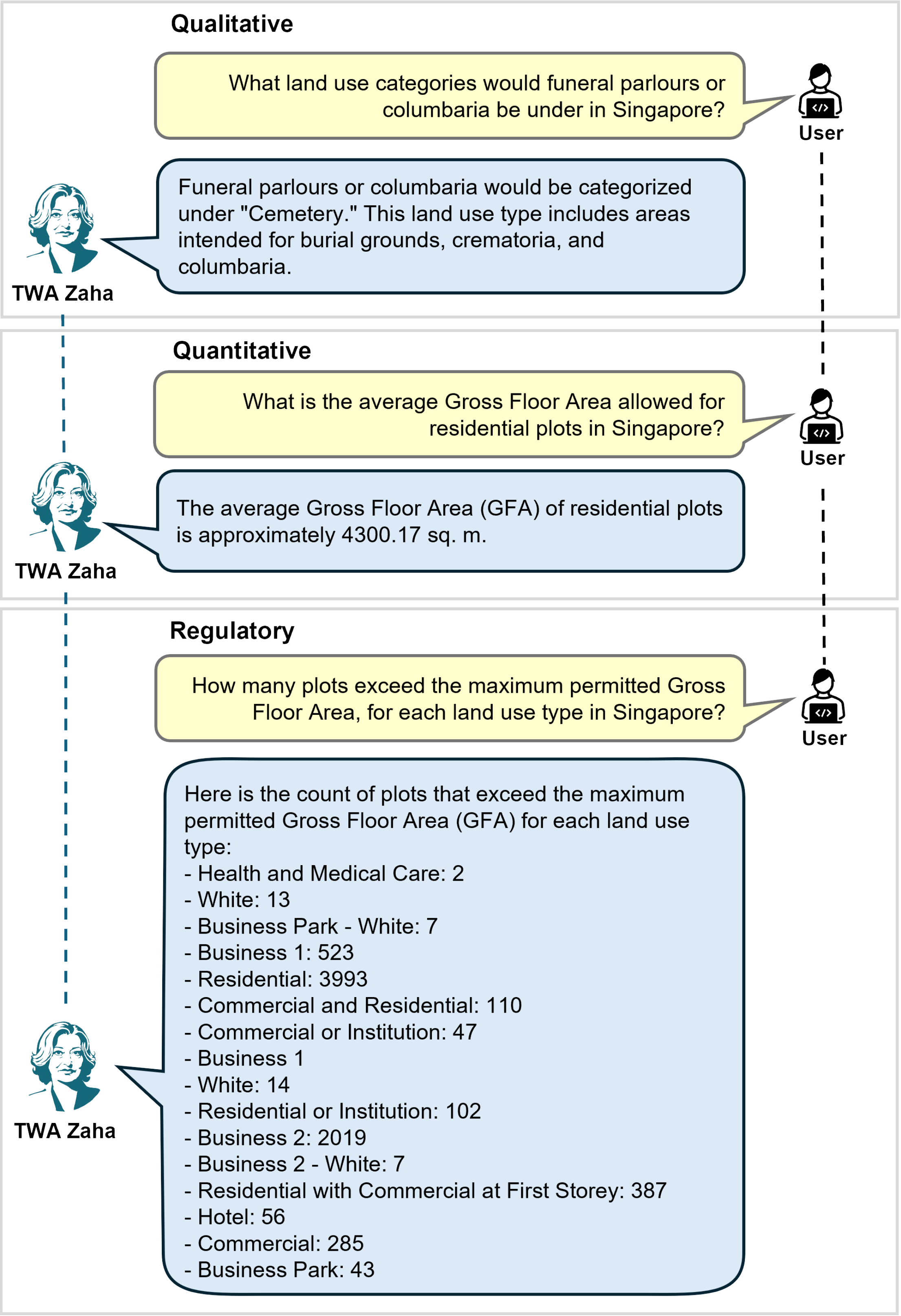
In all of these cases, Zaha predicts the SPARQL query required to access the knowledge graph, identifies the appropriate namespace, executes the query to retrieve the information, and then processes the results into natural language for ease of understanding. Through its integration with TWA knowledge graph, Zaha allows both experts and non-experts to easily access and interpret complex data, improving data management and usability in land use.

## 5.2 Real-time environmental monitoring

Zaha provides a natural language interface that simplifies access to air quality and dispersion data in TWA knowledge graph, making environmental monitoring data more interpretable for users without technical expertise.

**Figure** 6 shows an example of Zaha’s ability to query environmental data, where a user can ask questions such as, “What are the concentrations of air pollutants in Jurong Island, Singapore?”. Zaha processes this question by identifying relevant entities, such as “air pollutants” and “Jurong Island”, and connects these to the appropriate information sources in TWA knowledge graph.

To retrieve the necessary environmental data, Zaha makes a function call to a software



Qualitative

What land use categories would funeral parlours or columbaria be under in Singapore?

Funeral parlours or columbaria would be categorized under "Cemetery." This land use type includes areas intended for burial grounds, crematoria, and columbaria.

Quantitative

What is the average Gross Floor Area allowed for residential plots in Singapore?

The average Gross Floor Area (GFA) of residential plots is approximately 4300.17 sq. m.

Regulatory

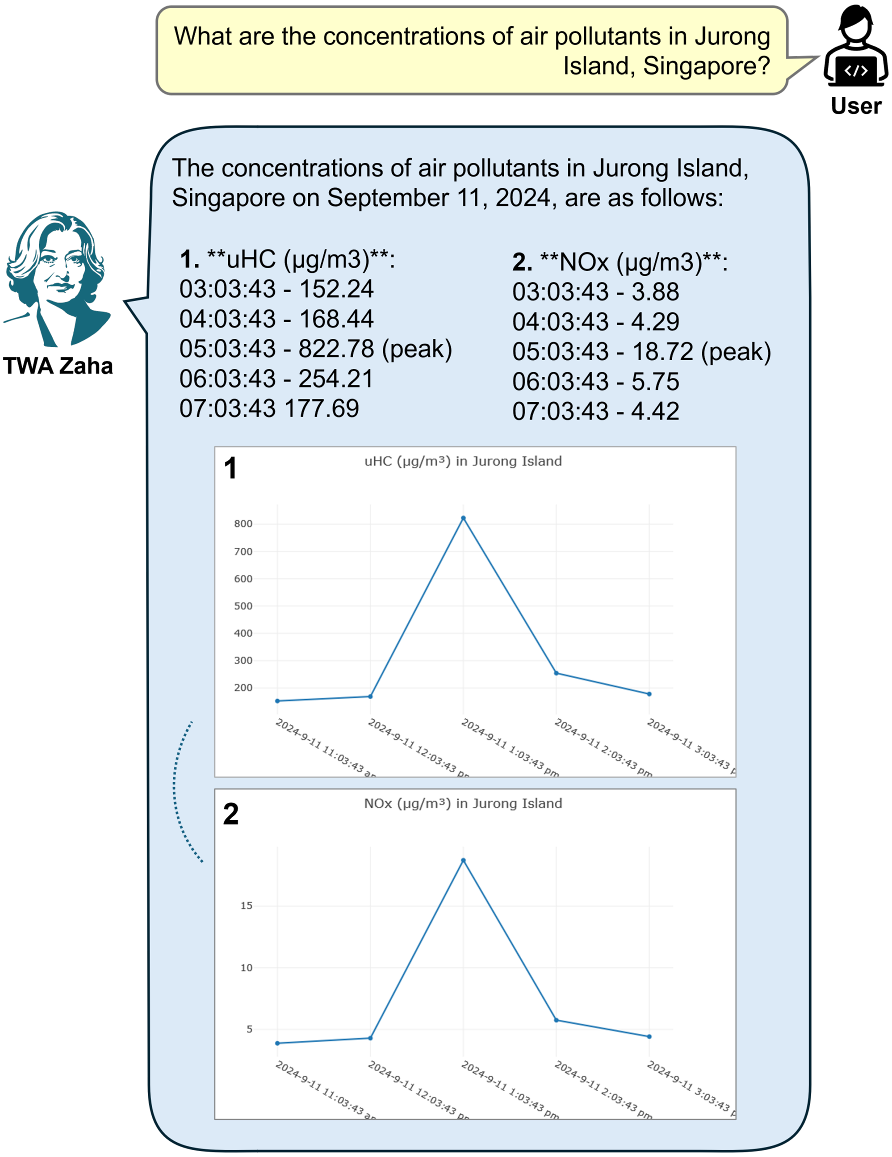
How many plots exceed the maximum permitted Gross Floor Area, for each land use type in Singapore?

Here is the count of plots that exceed the maximum permitted Gross Floor Area (GFA) for each land use type:
- Health and Medical Care: 2
- White: 13
- Business Park - White: 7
- Business 1: 523
- Residential: 3993
- Commercial and Residential: 110
- Commercial or Institution: 47
- Business 1
- White: 14
- Residential or Institution: 102
- Business 2: 2019
- Business 2 - White: 7
- Residential with Commercial at First Storey: 387
- Hotel: 56
- Commercial: 285
- Business Park: 43

**Figure 5:** *Zaha provides responses to various land use queries, demonstrating its ability to provide qualitative, quantitative, and regulatory insights.*

agent within TWA that queries recent time series on pollutant concentrations and ship activity around the user-specified area. When real-time dispersion data is unavailable, the agent employs AERMOD, a dispersion model by the U.S. Environmental Protection Agency [13], to simulate pollutant dispersion based on real-time conditions and ship traffic data. These simulations are continuously updated in the TWA knowledge graph to ensure relevance [25].

Zaha then presents this information as a concise natural language summary that highlights peak pollutant levels for the day, along with visualisations in the form of time series charts that show changes over time. This approach offers an intuitive interface for both expert and non-expert users, facilitating insights without the need for specialised knowledge of technical data formats or environmental models.



What are the concentrations of air pollutants in Jurong Island, Singapore?

The concentrations of air pollutants in Jurong Island, Singapore on September 11, 2024, are as follows:

1. **uHC (µg/m3)**:
03:03:43 - 152.24
04:03:43 - 168.44
05:03:43 - 822.78 (peak)
06:03:43 - 254.21
07:03:43 177.69

2. **NOx (µg/m3)**:
03:03:43 - 3.88
04:03:43 - 4.29
05:03:43 - 18.72 (peak)
06:03:43 - 5.75
07:03:43 - 4.42

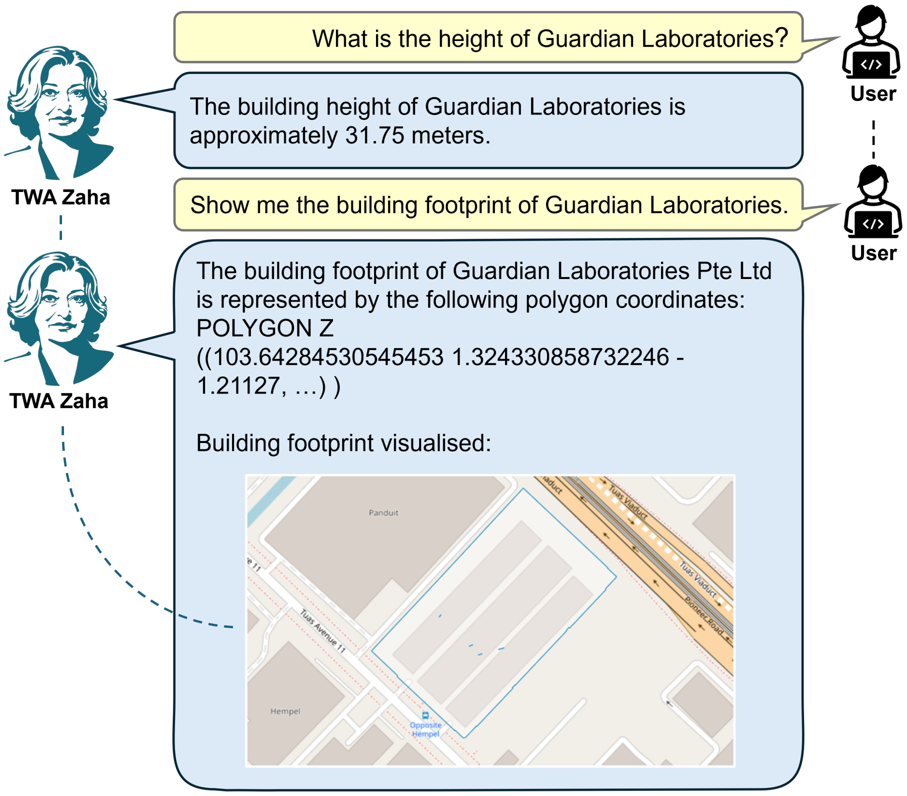
**Figure 6:** *Zaha provides natural language responses to air quality queries, detailing peak pollutant concentrations and graph-based visualisations for trend analysis.*

In addition to accessing air quality data, Zaha allows users to query live maritime information. For instance, if a user asks, “What is the speed over ground of the ship with MMSI 563071320?”, Zaha responds with a natural language description detailing the recorded ship speed and provides a chart that visualises speed variations over time. This real-time monitoring capability enhances decision-making in environmental management by allowing users to track maritime conditions as they evolve.

## 5.3 Integrated urban data

Zaha makes urban data easily accessible by allowing users to query building geometry and car park information with natural language, transforming complex data into clear, actionable insights. This capability streamlines access to urban data for all users, enhancing decision-making and simplifying the retrieval of essential urban information.

For example, as shown in **Fig.** 7 users can ask various questions regarding the geometry of the building, such as “What is the height of Guardian Laboratories?”, which Zaha can answer with natural language responses. Users can also further inquire, “Show me the building footprint of Guardian Laboratories”. In response, Zaha presents a map visualisation of the building’s footprint along with a Well-Known Text (WKT) string for GIS



What is the height of Guardian Laboratories?

The building height of Guardian Laboratories is approximately 31.75 meters.

Show me the building footprint of Guardian Laboratories.

The building footprint of Guardian Laboratories Pte Ltd is represented by the following polygon coordinates:
POLYGON Z
((103.64284530545453 1.324330858732246 -1.21127, ...) )

Building footprint visualised:

**Figure 7:** *Zaha provides natural language and visualised responses for building geometry queries, allowing users to assess building characteristics and spatial relationships effectively.*

applications. This dual response format, *i.e.,* textual and visual, facilitates quick identification of location and structure of the building, catering to both casual users and those needing precise GIS-compatible data.

Zaha also enables users to easily access information about car park availability through natural language queries. For instance, when a user asks, “Find me the car park nearest to CREATE Tower”, Zaha makes an external function call to retrieve the location of the closest car park, its distance from the specified building, and the current availability of parking spaces. This process enables Zaha to quickly provide essential parking information, enhancing accessibility for users managing urban spaces or seeking convenient parking options in densely populated areas.

# 6 Conclusions

In this work, we introduce Zaha, a knowledge graph-based question-answering system designed to streamline access to complex urban data through natural language queries. By leveraging The World Avatar dynamic knowledge graph, which provides a unified structure for integrating disparate data sources, Zaha effectively resolves issues related to urban data silos and inconsistent data formats. Furthermore, Zaha enables users to interact with complex urban data using natural language, making it accessible to both experts and non-experts, bypassing the technical barriers typically required to query the complex datasets.

Zaha leverages the power of RAG-based LLM system architecture and TWA dynamic knowledge graph to enhance retrieval accuracy and reduce hallucinations from generative AI. To further support Zaha’s querying capabilities, we have extended the ontological coverage of The World Avatar to describe car parks and regulations related to zones and land plots.

Currently, Zaha supports querying datasets related to land use planning, addressing both numerical and theoretical queries, as well as verifying rules and regulations. It can also access time series and generate graphs, which are particularly useful for monitoring environmental metrics. Additionally, Zaha enables users to obtain information about buildings, including their geometries and visualisations of building footprints, as well as to locate nearby car parks with real-time availability updates.

While Zaha demonstrates potential for natural language access for urban data, there is still room for improvement within the current architecture. Zaha relies on offline indexed data extracted from the TWA knowledge graph for accurate answer generation. To enhance integration with the dynamically changing TWA knowledge graph, automating the indexing process will allow Zaha to access the latest data while still leveraging the RAG architecture, which eliminates the need for retraining on new information. Additionally, enabling the capabilities of TWA agents as indexed inputs for the model can reduce the overhead associated with handcrafted datasets. This approach will also facilitate a more rapid expansion of Zaha’s functionalities within the TWA framework, rather than restricting it to the existing set of capabilities. Ultimately, the advancement of data-driven methodologies, exemplified by the development of Zaha, underscores the potential for enhanced urban planning and management, paving the way for more sustainable and resilient cities in the face of ongoing urbanisation challenges.

## 7 Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Markus Kraft gratefully acknowledges the support of the Alexander von Humboldt Foundation.

## Nomenclature

| | |
|---|---|
| **AI** | Artificial intelligence |
| **API** | Application programming interface |
| **BIM** | Building information modelling |
| **CSV** | Comma-Separated Values |
| **EPC** | Energy Performance Certificate |
| **GeoSPARQL** | Geographic query language for RDF Data |
| **GIS** | Geographic information system |
| **ICL** | In-context learning |
| **IFC** | Industry Foundation Classes |
| **IRI** | Internationalised resource identifier |
| **JSON** | JavaScript Object Notation |
| **KGQA** | Knowledge graph-based question answering |
| **KG** | Knowledge graph |
| **KMZ** | Keyhole Markup Language Zipped |
| **KNN** | K-nearest neighbours |
| **LLM** | Large language model |
| **NLP** | Natural language processing |
| **OSM** | Open street map |
| **PDF** | Portable Document Format |
| **QA** | Question answering |
| **RAG** | Retrieve-augmented generation |
| **RDF** | Resource Description Framework |
| **SPARQL** | SPARQL protocol and RDF query language |
| **TWA** | The World Avatar |
| **URA** | Singapore Urban Redevelopment Authority |
| **XML** | Extensible Markup Language |

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to enhance the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Data and code availability

All the codes developed are available on The World Avatar GitHub repository:
https://github.com/cambridge-cares/TheWorldAvatar.
Developed ontologies can be found in the ontology subdirectory and instructions to reproduce the use case are detailed in the

# A Appendix

## A.1 Namespaces

obe: <https://www.theworldavatar.com/kg/ontobuiltenv/>
ocp: <https://www.theworldavatar.com/kg/ocp/>
od: <https://www.theworldavatar.com/kg/ontodispersion/>
op: <https://www.theworldavatar.com/kg/ontoplot/>
opr: <https://www.theworldavatar.com/kg/ontoplotregulation/>
oz: <https://www.theworldavatar.com/kg/ontozoning/>
rdf: <http://www.w3.org/2000/01/rdf-schema#>
geo: <http://www.opengis.net/ont/geosparql#>
icontact: <http://ontology.eil.utoronto.ca/icontact.owl#>
ipark: <http://ontology.eil.utoronto.ca/icity/Parking/>

## A.2 System assessment test questions

The test questions used in Section 4.2 are shown in Table A.1.

**Table A.1:** *Test questions used for system assessment.*

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| What does Special Use mean as a land use category? | Complete entity name | Land use | Not applicable |
| What land use category would hospitals be classified as? | Semantically similar entity name | Land use | Not applicable |
| What's the difference between Business 1 and Business 2 land use classifications? | Multiple entities | Land use | Not applicable |
| How many land plots are designated for commercial use? | Semantically similar entity name | Land use | Not applicable |
| What is the percentage of plots zoned for residential use out of the total number of plots? | Semantically similar entity name | Land use | Not applicable |
| What is the area of the smallest agriculture lot? | Complete entity name | Land use | Not applicable |

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| What is the lowest gross plot ratio of lots designated for health and medical care facilities? | Complete entity name | Land use | Not applicable |
| Average GFA allowed for lots classified for business activities. | Semantically similar entity name | Land use | Not applicable |
| How many land lots do not exceed their max permitted GFA? | No entity in the question | Land use | Not applicable |
| What is the number of plots that exceed max permitted GFA for each land use type? | No entity in the question | Land use | Not applicable |
| List the number of plot for each land use type in Singapore. | No entity in the question | Land use | Not applicable |
| What is the land use size for Civic and Community Institution in Singapore? | Complete entity name | Land use | Not applicable |
| What is the number of plot for agricultural use in Singapore? | Semantically similar entity name | Land use | Not applicable |
| How many plots are designated for industrial purposes? | Semantically similar entity name | Land use | Not applicable |
| Find the number of plots with a maximum permitted GFA less than 250 square meters. | No entity in the question | Land use | Not applicable |
| What is the area of the smallest Reserve Site lot? | Complete entity name | Land use | Not applicable |
| What percentage of plots are designated for institutional purposes? | Semantically similar entity name | Land use | Not applicable |
| Which land use category has the least amount of land allocated? | No entity in the question | Land use | Not applicable |
| How many Commercial and Residential plots exceed their maximum permitted GFA? | Complete entity name | Land use | Not applicable |

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| How many industrial plots have a total gross floor area exceeding the permitted maximum? | Semantically similar entity name | Land use | Not applicable |
| What is the average gross floor area allowed for residential plots? | Complete entity name | Land use | Not applicable |
| What does "agricultural plot" mean in the land use classification? | Semantically similar entity name | Land use | Not applicable |
| How many plots are designated for mixed-use purposes? | Semantically similar entity name | Land use | Not applicable |
| How many plots designated for commercial use are awaiting detailed evaluation? | Complete entity name | Land use | Not applicable |
| What is the maximum gross plot ratio allowed for industrial facilities? | Semantically similar entity name | Land use | Not applicable |
| How many plots have a maximum permitted GFA between 300 and 700 square meters? | No entity in the question | Land use | Not applicable |
| What is the smallest lot size for residential plots? | Semantically similar entity name | Land use | Not applicable |
| What percentage of plots are designated for agriculture? | Complete entity name | Land use | Not applicable |
| What is the total gross floor area for Educational Institution plots? | Complete entity name | Land use | Not applicable |
| What is the maximum permitted GFA for healthcare facilities? | Semantically similar entity name | Land use | Not applicable |
| Compare the land use type "Institutional" and "Agricultural". | Multiple entities | Land use | Not applicable |
| How many land plots are designated for transport facility use? | Complete entity name | Land use | Not applicable |

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| How many agricultural plots and hotel plots are awaiting detailed evaluation? | Multiple entities | Land use | Not applicable |
| How many institutional plots with a GFA over 600 square meters have an area greater than 1000 square meters? | Semantically similar entity name | Land use | Not applicable |
| How many commercial land lots do not exceed their max permitted GFA? | Semantically similar entity name | Land use | Not applicable |
| What is the GFA range for commercial buildings? | Semantically similar entity name | Land use | Not applicable |
| Compare the land use type "Commercial" and "Residential". | Multiple entities | Land use | Not applicable |
| What is the maximum GFA for plots in commercial areas with an area exceeding 1500 square meters? | Semantically similar entity name | Land use | Not applicable |
| What is the smallest plot area for residential plots with a GFA exceeding 1200 square meters? | Semantically similar entity name | Land use | Not applicable |
| How many plots classified as "Commercial" have an area of over 1000 square meters? | Semantically similar entity name | Land use | Not applicable |
| In total, how many buildings are there in Singapore? | No entity in the question | Urban data | Not applicable |
| What's the building height of APEC INDUSTRIES? | Complete entity name | Urban data | Not applicable |
| What's the building height of APEC INDSTRIES? | Entity name with typographical errors | Urban data | Not applicable |
| Visualise the building footprint of Abbott Manufacturing. | Complete entity name | Urban data | Not applicable |
| Visualise the building footprint of Abbott Mnaufcaturing. | Entity name with typographical errors | Urban data | Not applicable |
| How many office buildings are there? | No entity in the question | Urban data | Not applicable |

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| What are the 5 most common building usages? | No entity in the question | Urban data | Not applicable |
| What is the tallest building height in Singapore? | No entity in the question | Urban data | Not applicable |
| What is the average height of industrial buildings in Singapore? | No entity in the question | Urban data | Not applicable |
| How many residential buildings are there in Singapore? | No entity in the question | Urban data | Not applicable |
| How many hospital buildings are there? | No entity in the question | Urban data | Not applicable |
| What are the 3 most common building purposes? | No entity in the question | Urban data | Not applicable |
| What is the height of the ST Engineering Hub? | Complete entity name | Urban data | Not applicable |
| What is the height of the Sst Engineering Hub? | Entity name with typographical errors | Urban data | Not applicable |
| What is the maximum calculated GFA for industrial buildings? | No entity in the question | Urban data | Not applicable |
| Show the footprint of New Tech Park. | Complete entity name | Urban data | Not applicable |
| Show the footprint of New Tech Pak. | Entity name with typographical errors | Urban data | Not applicable |
| How many buildings are taller than 50 meters? | No entity in the question | Urban data | Not applicable |
| What is the average GFA for hotel buildings? | No entity in the question | Urban data | Not applicable |
| What is the tallest office building in Singapore? | No entity in the question | Urban data | Not applicable |
| What are the top 5 building usages by count? | No entity in the question | Urban data | Not applicable |
| What is the building height of Visa Singapore data centre? | Complete entity name | Urban data | Not applicable |
| What is the building height of Viaa Singapore data centre? | Entity name with typographical errors | Urban data | Not applicable |
| What is the lowest calculated GFA among retail buildings? | No entity in the question | Urban data | Not applicable |

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| What is the average height of school buildings? | No entity in the question | Urban data | Not applicable |
| What is the total area covered by hospital buildings in Singapore? | No entity in the question | Urban data | Not applicable |
| What is the average GFA of industrial buildings in Singapore? | No entity in the question | Urban data | Not applicable |
| What is the highest GFA for data centres in Singapore? | No entity in the question | Urban data | Not applicable |
| How many buildings are taller than 100 meters in Singapore? | No entity in the question | Urban data | Not applicable |
| What is the maximum calculated GFA for government buildings? | No entity in the question | Urban data | Not applicable |
| What are the concentrations of air pollutants in Jurong Island, Singapore? | Not applicable | Not applicable | Air pollution |
| What are the air pollutant concentration readings in CREATE building, Singapore? | Not applicable | Not applicable | Air pollution |
| What are the concentrations of air pollutants in Utown Singapore? | Not applicable | Not applicable | Air pollution |
| What are the concentrations of air pollutants in NUS Central Library Singapore? | Not applicable | Not applicable | Air pollution |
| What are the concentrations of air pollutants in EXXON PAC Singapore? | Not applicable | Not applicable | Air pollution |
| what's the speed over ground of ship with MMSI 563071320? | Complete entity name | Not applicable | Ship information |
| What's the maximum static draught of ship MMSI 563071330? | Complete entity name | Not applicable | Ship information |
| Get all the information for MMSI 565102000. | Complete entity name | Not applicable | Ship information |
| What is the IMO number of ship identified by MMSI 564743000? | Complete entity name | Not applicable | Ship information |

| Question | Entity linking scenario | SPARQL generation topic | Function call topic |
|---|---|---|---|
| Plot the variation in course over ground of ship MMSI 565102000. | Complete entity name | Not applicable | Ship information |
| Find me the carpark nearest to Vivo City Singapore. | Not applicable | Not applicable | Carpark information |
| Find me the carpark nearest to Marina Bay Sand Singapore. | Not applicable | Not applicable | Carpark information |
| Find me the carpark nearest to Suntec City Singapore. | Not applicable | Not applicable | Carpark information |
| Find me the carpark nearest to Jurong Island, Singapore. | Not applicable | Not applicable | Carpark information |
| Find me the carpark nearest to CREATE tower, Singapore. | Not applicable | Not applicable | Carpark information |
| Find me the carpark nearest to Changi Airport Singapore. | Not applicable | Not applicable | Carpark information |

# References

[1] Open Data DC — opendata.dc.gov. https://opendata.dc.gov/. [Accessed 02-10-2024].

[2] J. Akroyd, S. Mosbach, A. Bhave, and M. Kraft. Universal Digital Twin - A Dynamic Knowledge Graph. *Data-Centric Engineering*, 2:e14, 2021. doi:10.1017/dce.2021.10.

[3] A. Arbaaeen and A. Shah. Ontology-Based Approach to Semantically Enhanced Question Answering for Closed Domain: A Review. *Information*, 12(5):200, May 2021. doi:10.3390/info12050200.

[4] L. Barbosa, K. Pham, C. Silva, M. R. Vieira, and J. Freire. Structured Open Urban Data: Understanding the Landscape. *Big Data*, 2(3):144–154, Sept. 2014. doi:10.1089/big.2014.0020.

[5] H. Bast, B. Buchhold, and E. Haussmann. Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, 10(2-3):119–271, 2016.

[6] M. Batty. Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3):274–279, Nov. 2013. doi:10.1177/2043820613513390.

[7] L. M. Bettencourt. The Uses of Big Data in Cities. *Big Data*, 2(1):12–22, Mar. 2014. doi:10.1089/big.2013.0042.

[8] S. E. Bibri. *Unprecedented Innovations in Sustainable Urban Planning: Novel Analytical Solutions and Data-Driven Decision-Making Processes*, page 247–296. Springer International Publishing, 2018. ISBN 9783319739816. doi:10.1007/978-3-319-73981-6_5.

[9] S. E. Bibri and J. Krogstie. The emerging data–driven Smart City and its innovative applied solutions for sustainability: the cases of London and Barcelona. *Energy Informatics*, 3(1), June 2020. doi:10.1186/s42162-020-00108-6.

[10] Blazegraph. Blazegraph, 2020. URL https://blazegraph.com. Accessed: 25 September 2024.

[11] D. Calì, A. Condorelli, S. Papa, M. Rata, and L. Zagarella. Improving intelligence through use of Natural Language Processing. A comparison between NLP interfaces and traditional visual GIS interfaces. *Procedia Computer Science*, 5:920–925, 2011.

[12] M. A. Calijorne Soares and F. S. Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, July 2020. doi:10.1016/j.jksuci.2018.08.005.

[13] A. J. Cimorelli, S. G. Perry, A. Venkatram, J. C. Weil, R. Paine, R. B. Wilson, R. F. Lee, W. D. Peters, and R. W. Brode. AERMOD: A Dispersion Model for Industrial Source Applications. Part I: General Model Formulation and Boundary

Layer Characterization. *Journal of Applied Meteorology*, 44(5):682–693, May 2005. doi:10.1175/jam2227.1.

[14] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[15] DC Compass. Compass — opendata.dc.gov. https://opendata.dc.gov/pages/compass. [Accessed 02-10-2024].

[16] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, and Z. Sui. A Survey on In-context Learning, 2023. URL https://arxiv.org/abs/2301.00234.

[17] M. Duerst and M. Suignard. *Internationalized Resource Identifiers (IRIs)*. 2005. doi:10.17487/rfc3987.

[18] Eclipse Foundation. Eclipse RDF4J, 2024. URL https://rdf4j.org/. Accessed: 25 September 2024.

[19] Z. Engin, J. van Dijk, T. Lan, P. A. Longley, P. Treleaven, M. Batty, and A. Penn. Data-driven urban management: Mapping the landscape. *Journal of Urban Management*, 9(2):140–150, June 2020. doi:10.1016/j.jum.2019.12.001.

[20] Y. Feng, L. Ding, and G. Xiao. GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base (Short Paper). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPICS.GISCIENCE.2023.28. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.GIScience.2023.28.

[21] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. doi:10.1006/knac.1993.1008.

[22] D. Guo, E. Onstein, and A. D. L. Rosa. An Approach of Automatic SPARQL Generation for BIM Data Extraction. *Applied Sciences*, 10(24):8794, Dec. 2020. doi:10.3390/app10248794.

[23] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, July 2015. doi:10.1126/science.aaa8685.

[24] M. Hofmeister, J. Bai, G. Brownbridge, S. Mosbach, K. F. Lee, F. Farazi, M. Hillman, M. Agarwal, S. Ganguly, J. Akroyd, and M. Kraft. Semantic agent framework for automated flood assessment using dynamic knowledge graphs. *Data-Centric Engineering*, 5, 2024. doi:10.1017/dce.2024.11.

[25] M. Hofmeister, K. F. Lee, Y.-K. Tsai, M. Müller, K. Nagarajan, S. Mosbach, J. Akroyd, and M. Kraft. Dynamic control of district heating networks with integrated emission modelling: A dynamic knowledge graph approach. *Energy and AI*, 17:100376, Sept. 2024. doi:10.1016/j.egyai.2024.100376.

[26] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, 2023. URL https://arxiv.org/abs/2311.05232.

[27] X. Huang, J. Zhang, D. Li, and P. Li. Knowledge Graph Embedding Based Question Answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, volume 29 of *WSDM '19*, page 105–113. ACM, Jan. 2019. doi:10.1145/3289600.3290956.

[28] Hugging Face. sentence-transformers/all-mpnet-base-v2 · Hugging Face — huggingface.co, 2021. URL https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed: 18 October 2024.

[29] M. Katsumi. iCity Contact Ontology — enterpriseintegrationlab.github.io. https://enterpriseintegrationlab.github.io/icity/Contact/doc/index-en.html, . [Accessed 25-09-2024].

[30] M. Katsumi. iCity Parking Ontology — enterpriseintegrationlab.github.io. https://enterpriseintegrationlab.github.io/icity/Parking/doc/index-en.html, . [Accessed 25-09-2024].

[31] I. Kistelegdi, K. R. Horváth, T. Storcz, and Z. Ercsey. Building Geometry as a Variable in Energy, Comfort, and Environmental Design Optimization—A Review from the Perspective of Architects. *Buildings*, 12(1):69, Jan. 2022. doi:10.3390/buildings12010069.

[32] A. Kondinski, P. Rutkevych, L. Pascazio, D. N. Tran, F. Farazi, S. Ganguly, and M. Kraft. Knowledge graph representation of zeolitic crystalline materials. *Digital Discovery*, 2024. doi:10.1039/d4dd00166d.

[33] M. Kraft and S. Mosbach. The future of computational modelling in reaction engineering. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1924):3633–3644, Aug. 2010. doi:10.1098/rsta.2010.0124.

[34] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10(8):708–710, 1966.

[35] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[36] Z. Li and H. Ning. Autonomous GIS: the next-generation AI-powered GIS. *International Journal of Digital Earth*, 16(2):4668–4686, Nov. 2023. doi:10.1080/17538947.2023.2278895.

[37] M. Q. Lim, X. Wang, O. Inderwildi, and M. Kraft. *The World Avatar—A World Model for Facilitating Interoperability*, page 39–53. Springer International Publishing, 2022. ISBN 9783030862152. doi:10.1007/978-3-030-86215-2_4.

[38] A. MacLachlan, E. Biggs, G. Roberts, and B. Boruff. Sustainable City Planning: A Data-Driven Approach for Mitigating Urban Heat. *Frontiers in Built Environment*, 6, Jan. 2021. doi:10.3389/fbuil.2020.519599.

[39] A. Mukanova, M. Milosz, A. Dauletkaliyeva, A. Nazyrova, G. Yelibayeva, D. Kuzin, and L. Kussepova. LLM-Powered Natural Language Text Processing for Ontology Enrichment. *Applied Sciences*, 14(13):5860, July 2024. doi:10.3390/app14135860.

[40] MyCity Chatbot. MyCity Chatbot — chat.nyc.gov. https://chat.nyc.gov/. [Accessed 02-10-2024].

[41] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1):31–88, 2001.

[42] OpenAI. GPT-4o mini: advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 18 October 2024.

[43] L. Pascazio, D. Tran, S. D. Rihm, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. Question-answering system for combustion kinetics. *Proceedings of the Combustion Institute*, 40(1–4):105428, 2024. doi:10.1016/j.proci.2024.105428.

[44] J. C. Rangel, T. M. de Farias, A. C. Sima, and N. Kobayashi. SPARQL Generation: an analysis on fine-tuning OpenLLaMA for Question Answering over a Life Science Knowledge Graph, 2024. URL https://arxiv.org/abs/2402.04627.

[45] Redis. Redis - The Real-time Data Platform — redis.io. https://redis.io/, 2024. Accessed: 18 October 2024.

[46] S. Setty, H. Thakkar, A. Lee, E. Chung, and N. Vidra. Improving Retrieval for RAG based Question Answering Models on Financial Documents, 2024. URL https://arxiv.org/abs/2404.07221.

[47] A. Sheth, S. Padhee, and A. Gyrard. Knowledge Graphs and Knowledge Networks: The Story in Brief, 2020. URL https://arxiv.org/abs/2003.03623.

[48] B. N. Silva, M. Khan, C. Jung, J. Seo, D. Muhammad, J. Han, Y. Yoon, and K. Han. Urban Planning and Smart City Decision Management Empowered by Real-Time Data Processing Using Big Data Analytics. *Sensors*, 18(9):2994, Sept. 2018. doi:10.3390/s18092994.

[49] H. Silvennoinen, A. Chadzynski, F. Farazi, A. Grišiūtė, Z. Shi, A. von Richthofen, S. Cairns, M. Kraft, M. Raubal, and P. Herthogs. A semantic web approach to land use regulations in urban planning: The ontozoning ontology of zones, land uses and programmes for Singapore. *Journal of Urban Management*, 12(2):151–167, June 2023. doi:10.1016/j.jum.2023.02.002.

[50] E. J. Taylor and R. van Bemmel-Misrachi. The elephant in the scheme: Planning for and around car parking in Melbourne, 1929–2016. *Land Use Policy*, 60:287–297, Jan. 2017. doi:10.1016/j.landusepol.2016.10.044.

[51] P. Thakuriah, N. Y. Tilahun, and M. Zellner. *Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery*, page 11–45. Springer International Publishing, Oct. 2016. ISBN 9783319409023. doi:10.1007/978-3-319-40902-3_2.

[52] D. N. Tran, S. D. Rihm, A. Kondinski, L. Pascazio, F. Saluz, S. Mosbach, J. Akroyd, and M. Kraft. Natural Language Access Point to Digital Metal-Organic Polyhedra Chemistry in The World Avatar, 2024. Submitted for publication. Preprint available at https://como.ceb.cam.ac.uk/preprints/327/.

[53] Urban Redevelopment Authority. Updates to the landscaping for urban spaces and high-rises (LUSH) programme: LUSH 3.0. https://www.ura.gov.sg/Corporate/Guidelines/Circulars/dc17-06, 2017. [Accessed 22-10-2024].

[54] Urban Redevelopment Authority. Gross Floor Area. https://www.ura.gov.sg/Corporate/Guidelines/Development-Control/gross-floor-area/GFA/Introduction, 2023. [Accessed 22-10-2024].

[55] M. Uschold and M. Gruninger. Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93–136, June 1996. doi:10.1017/s0269888900007797.

[56] E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR00. ACM, July 2000. doi:10.1145/345508.345577.

[57] W3C. SPARQL 1.1 Query Language, 2013. URL https://www.w3.org/TR/sparql11-query/. Accessed: 25 September 2024.

[58] N. Wang, R. R. A. Issa, and C. J. Anumba. NLP-Based Query-Answering System for Information Extraction from Building Information Models. *Journal of Computing in Civil Engineering*, 36(3), May 2022. doi:10.1061/(asce)cp.1943-5487.0001019.

[59] G. Xiao, D. Lanti, R. Kontchakov, S. Komla-Ebri, E. Güzel-Kalaycı, L. Ding, J. Corman, B. Cogrel, D. Calvanese, and E. Botoeva. The virtual knowledge graph system ontop. In *International Semantic Web Conference*, pages 259–277. Springer, 2020. doi:10.1007/978-3-030-62466-8_17.

[60] A. G. Yeh. Urban planning and GIS. *Geographical information systems*, 2(877-888): 1, 1999.

[61] Y. Zhang, H. Dai, Z. Kozareva, A. Smola, and L. Song. Variational Reasoning for Question Answering With Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi:10.1609/aaai.v32i1.12057.

[62] B. Zhong, W. He, Z. Huang, P. E. Love, J. Tang, and H. Luo. A building regulation question answering system: A deep learning methodology. *Advanced Engineering Informatics*, 46:101195, Oct. 2020. doi:10.1016/j.aei.2020.101195.

[63] X. Zhou, A. Eibeck, M. Q. Lim, N. B. Krdzavac, and M. Kraft. An agent composition framework for the J-Park Simulator - A knowledge graph for the process industry. *Computers & Chemical Engineering*, 130:106577, 2019. doi:10.1016/j.compchemeng.2019.106577.

[64] X. Zhou, S. Zhang, M. Agarwal, J. Akroyd, S. Mosbach, and M. Kraft. Marie and BERT-A Knowledge Graph Embedding Based Question Answering System for Chemistry. *ACS Omega*, 8(36):33039–33057, Aug. 2023. doi:10.1021/acsomega.3c05114.