Improving Tourism Analytics From Climate Data Using Knowledge Graphs

Jiantao Wu [ORCID], Jarrett Pierse [ORCID], Fabrizio Orlandi, Declan O'Sullivan [ORCID], and Soumyabrata Dev [ORCID], *Member, IEEE*

Abstract—**Climate change has been deemed to be one of the greatest challenges facing humans in the 21st century, with extreme weather events taking place more regularly than before. While the impact of climate change has been well documented in recent years across industries, the impact of climate change on the tourism economy is yet to be fully realized. This article aims to apply a range of knowledge graph techniques to naturalistic data. Among these, weather data will be explored as one prospective way to enhance people's understanding of how climate and a country's tourism economy are related and how they interact. According to our exploration with the knowledge graph approach in organizing the climate and tourism data, the insights and knowledge gained from the knowledge graph are able to ultimately help improve the quality of life for people and the tourism industry of a country.**

Index Terms—**Climate–tourism analytics, flight data, knowledge graph, linked data.**

I. INTRODUCTION

CLIMATE change is one of the most well-recognized and published areas of scientific research over the last three decades. Oo et al. [1] assessed climate change projection models on a global scale with an analysis of increased precipitation, average global temperatures, and the levels of atmospheric gases associated with human activities. According to the assessment result, the projections for changes in average monthly maximum temperatures increase in the Upper Ayeyarwady river basin under three climate representative concentration pathway (RCP) indices: RCP2.6, RCP4.5, and RCP8.5. A 2019 statistical study [2] of peer-reviewed scientific literature surrounding climate change found a 99% consensus on the existence of climate change. Further research has documented the relationship between climate change and tourism, such as in this simulation study [3], where projections showed that tourism trends might move toward higher elevations and latitudes due to climate change. Further impacts from climate change are concluded in this article [4] that the tourism industry throughout the world will likely be severely impacted by climate change. Particularly notable are the regional vulnerabilities that vary. Small island governments and emerging areas in Africa, Asia, and Oceania look to be at the most danger as a result of prospective demand changes preferring higher latitude nations. Given these facts, it is evident that climate change will have far-reaching impacts on the worldwide tourism industry. In addition, climatic parameters [5], such as precipitation, altitude, and extreme weather events, have been proven to have a direct association with tourism demand. Li et al. [6] explored the links between climate and tourism via three statistical models of tourist data between Hong Kong and Mainland China. The results found that destination, home, and source market climates all affect Hong Kong residents' appetite for travel in Mainland China. This indicates that "as the climate difference between home and destination climates increases, so does the likelihood of tourism from that home region to that destination." The results further conclude that *maximum daily temperature* had the strongest positive impact on tourism when compared with a total of six daily climatic variables: maximum temperature, minimum temperature, average temperature, average humidity, average precipitation, and average hours of sunshine.

On the other hand, a knowledge graph connects nodes based on a semantic relationship with minimal bounds [7]. Data points can be connected at multiple levels of abstraction within the same graph, and crucially, new relations can be derived from existing data [8]. This proves to be extremely powerful in modeling real-world entities and relationships due to how the knowledge graph (KG) captures semantic meaning, making it useful for obtaining intelligence through artificial intelligence and machine learning [9], [10], [11]. The climate–tourism domain is data-rich, with a vast amount of data generated every day. Knowledge graphs offer benefits in relating vast amounts of data but are underutilized in this domain.

Given the connection between climate and tourism in the studies discussed above and the fact that the advantages of using advanced knowledge graph techniques are underexplored for improved climate and tourism data analysis, this article [12] proposes a "Knowledge graph-based Tourism Analytic System" where we explore how knowledge graph techniques, such as RDF and SPARQL queries, can efficiently organize climate and tourism data in a way to provide users with better location,

lodging, and local attraction insights. Specifically, the climate data are collected from North America, including rainfall, temperature, and wind data, and will be integrated with the tourism data, such as flight data, to create a multidomain ontology.

A. Structure of This Article

The rest of this article is organized as follows. Section II discusses the relevant literature over which the climate–tourism knowledge graph is proposed to fill the current research gap. Section III describes the technique for constructing a climate–tourist KG from the schema and architectural design of the KG as well as populating the KG platform with real-world climate and tourism data. In Section IV, we evaluate the KG platform using a standard usability study and explain the results in terms of the usability of the system. In addition, we demonstrate the application value of the proposed climate–tourism KG by providing examples of common use cases. We explore some limits of existing work in Section VI. Finally, Section VII concludes this article.

B. Contributions of This Article

1) We provide a knowledge graph schema for organizing diverse sources of climate data and tourist data that are reusable and adaptive to schema modifications.
2) To the best of our knowledge, we first integrated various real-world data sources of climate observations [e.g., National Oceanic and Atmospheric Administration (NOAA)], flight data (e.g., AviationStack), map data (e.g., SimpleMaps), etc. to create a climate–tourism KG platform that improves tourism analytics by taking climate into consideration.
3) We provide systematic insights into the effectiveness of the proposed climate–tourism KG platform by combining usability test, use case study, and a discussion of limitations.
4) The proposed climate–tourism knowledge graph and its semantic schema have been published on the Web,[1] which can be potentially reused to allow schema sharing across heterogeneous data sources via knowledge graph techniques, such as Linked Data principles [13].

II. RELATED WORK

In this section, we discuss the related work on some existing challenges in building a knowledge graph containing multi-source data, such as data ambiguity, as well as the insights that inspire our work from some sophisticated industry-level knowledge graphs that are already in use for clients, as there are currently few sophisticated knowledge graphs built at the intersection of climate and tourism.

A. Multisource Data Knowledge Graph Development

While the rise of knowledge-based data and KGs has increased over the past decade, there remain some significant

---

[1]In the spirit of reproducible research, all the source code is available at https://github.com/futaoo/climate-tourism-kg.

challenges in developing intelligent knowledge graphs. Cizer et al. [13] outlined some of the significant challenges, primarily the processing of data mapping (schema mapping) and data fusion. Once data are obtained from various sources, it must be integrated before being presented or processed. There have been improvements in recent years in how data are dynamically fused to create larger ontologies through the use of algorithms.

Novak et al. [14] conducted a large study of over 600 participants across seven European cities. This study involved collecting sensor data on local air quality and integrating these sensor data together to provide residents with a comprehensive view of the air quality in their local vicinity. With a large variation in the types of sensor devices used and the data they produced, there was a need to fuse these data together in a way that prevented gaps in knowledge. On one device, there was no real-time clock built-in, and thus, the data needed to be timestamped correctly to produce a clear timeline. To achieve this, they utilized an online server and SIM card to log the time; however, this remained a challenge due to irregular and unreliable timestamp logging.

Data fusion is not the only challenge facing KGs. Precision and accuracy of facts remains a challenge when collecting a large sample of data. Wiekum [15] further explained how increased coverage leads to decreased correctness. To address this challenge, the decision is made to prioritize accuracy over recall (coverage) by building a knowledge graph of near-human quality with error rates comparable to expert curation, around 1%–5%. Furthermore, resolving data conflicts proves to be a challenge across all knowledge-based systems.

In the studies [13], [14], and [15], the challenge of resolving ambiguity when multiple sources provide different values for the same real-world entity is addressed. Solutions to this challenge in traditional databases have been utilized to resolve these data conflicts. In another paper [16], a *multifusion similarity map* was used to compare molecular compounds from multiple databases against a set of generalized compounds to search for similarity. A solution to this challenge in KGs is to canonicalize all entities so that they are uniquely identifiable. Typically, a canonicalized representation captures the alias names for each entity and groups statements per entity, not per name [15]. This prevents inconsistencies across the KGs. For instance, Wu et al. [17] proposed using a semantic approach to organize and model decentralized energy networks, where each energy metering result is characterized as a URI-encoded entity instance of the unambitious class "ElectricPowerEvaluation" specified by the SEAS ontology [18].

B. Lessons for Climate and Tourism Knowledge Graph

Given that this work uses climate data with knowledge graphs, it must be acknowledged that reorganizing climate data can be a significant challenge, particularly when aiming to create intelligence on these datasets [19], [20]. Due to this challenge, additional efforts must be made to ensure that the integrated ontology consists of high-quality data and verified sources. Due to its prominence in the climate research community, NOAA climate data are often used to generate climate-centric knowledge graphs [21].

To the best of our knowledge, there are currently a few sophisticated KGs use cases in the application fields of climate and tourist data. Nonetheless, there are some valuable lessons to be learned from well-known and established KGs, such as Google's [22], AceKG [23] and Diffbot [24]. Google is commonly credited with popularizing the term "Knowledge Graph," despite the underlying concepts and term existing for many years prior to 2012. The primary goal of the Google KG is to simplify the search process by providing users with more relevant facts and information related to a search query. According to an updated blog post from Google in May 2020; the Google KG now consists of over 500 billion facts about five billion entities [25]. This feat of engineering demonstrates the practicability of KGs as well as the reliance many of us have on them, often without notice.

The AceKG [23] knowledge graph provides over 3.13 billion triples surrounding academic facts, including research papers, authors, domains, organizations, and institutions. The usefulness of a KG such as AceKG is clear when considering how scientific research is conducted. Google Scholar[2] is widely used when researching papers and scientific journals online, although there is little public information on the mechanisms through which it does this. Given the prevalence of KGs in the Google search engine, it is likely that similar technologies are utilized by Google Scholar.

Diffbot [26] provides businesses and organizations with enterprise solutions that utilize their Web API. This includes exploiting KG technologies to search for organizations, news, retail products, and events. As funding for AI-powered services and tools continues to rise, Diffbot is one of many competitors in this space.

In this work, the aim is to apply comparable knowledge graph technologies to climate and tourism data. This domain of knowledge is relatively unexplored with a larger focus of KGs on enterprise and search engine solutions. Despite this focus, the underlying concepts of knowledge graphs, as mentioned before in Section I, are theoretically applicable to a diverse range of problems and datasets, with varying difficulty and challenges yet unrealized potential. Thus, the proposal of building a KG in the climate–tourism domain is presented.

III. METHODOLOGY

A. Schema and Architecture Design

The schema of a KG is often defined in accordance with particular domain applications. This study initially considers the field of application of KG to provide weather information to popular air travel services. The project schema, shown in Fig. 1, provides a comprehensive overview of the entities and relations in the KG. We manually defined this schema according to some necessary scenarios that exploit weather information for tourism. For instance, considering the scenario that users may be interested in an airport's local weather, we provided the KG with "NEARBY" (see Fig. 1) semantics for querying the nearby available weather observations for any airport. The
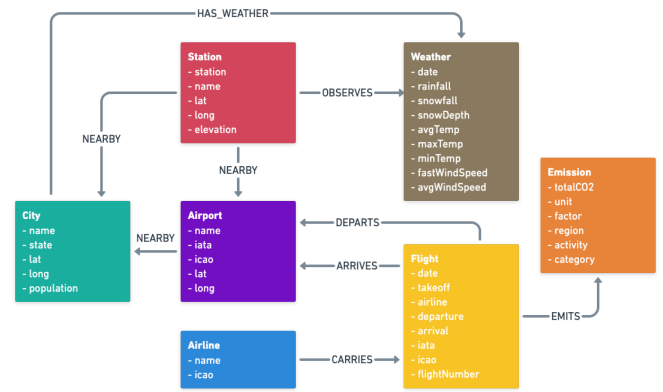

Fig. 1. Proposed schema.

overall schema definition was done using the CYPHER query language, which is natively supported by Neo4j, and the overall process can be described intuitively as follows.[3]

Initially, entities were classified based on real-world objects such as *City* and *Station*. As the project developed, entities such as *Emission* and *Weather* were added to describe the more abstract real-world concepts of CO_2 emissions and a daily weather event. Entities were initially chosen where a clear relationship between them existed, such as *Station* - NEARBY -> *City*. This was later expanded to include multiple NEARBY relationships: *Airport* - NEARBY -> *City* and *Station* - NEARBY -> *Airport*. This creates a three-way relationship between these entities, meaning any two entities can be queried without the need for the third. Initially, the KG connected *Station* to *Airport* through a path such as *Station* - NEARBY - *City* - NEARBY -> *Airport*. By adding the additional relation *Station* - NEARBY -> *Airport*, this type of path can be avoided when executing queries. The *City* - HAS_WEATHER -> *Weather* relation was added at a later stage to describe the abstract relation between *City* and *Weather* as initially weather entities were only connected to the KG via the "OBSERVES" relation. The addition of this relation improves the usability of the graph and more accurately reflects the real world.

Fig. 2 provides an overview of the architecture components. Data were collected from multiple sources, and separate data parsers were developed to transform the data from various forms into a standardized format for ingestion into the Neo4j DBMS. Weather data were appended to a climate dataset for each weather station in the United States. Airport and flight data were parsed in `JSON` format and converted into a `CSV` file. Emissions estimations were returned in `JSON` format and converted to `CSV` as appropriate. The Cities dataset was a static `CSV` file, taken from SimpleMaps.[4] The relationships between entities were defined through intermediary CSV files containing a reference to each entity; for example, *Station* and *City*. These CSV files were then ingested into Neo4j DBMS through the

---

[2]https://scholar.google.com/

[3]Each entity contains the relevant node properties associated with a node of that type; for example, the *Station* entity has the node property *lat*, the value of the latitude location of a weather station. The edges describe the relations between entities, such as "CARRIES."

[4]https://simplemaps.com

Fig. 2. Overview of the proposed architecture.

CYPHER query language to build the KG. Further data can be integrated into the KG through the addition of new entities and relations, i.e., flexible schema modification in Fig. 1. A prospective entity and relation that was not explored in this project includes integrating *Hotel* and *Booking* entities that could be extracted from a booking platform, such as Booking.[5]

B. Climate–Tourism Knowledge Graph Construction

1) Data Collection: Since the rise of the Semantic Web in the late 1990s, various methods of data collection have been applied to connect datasets across the Internet. Nowadays, the way data are distributed over the Internet has become increasingly standardized, with datasets being published on open web servers, making it easier to connect Linked Data [27] together. The book [28] outlines modern methods of data collection using standards and tools such as RDF, OWL, and SPARQL. Khan Academy [29] is an example of an online open source learning platform that utilizes KGs as part of their *learning recommendation system*; however, these KGs are usually manually constructed by experienced domain experts [30], for example, the teachers. This method of collecting and organizing data is not well equipped to scale as the KG coverage increases and additional entities and relations are added; thus, an automated solution must be utilized for large KG construction. Chen et al. [30] proposed a system to automatically construct KGs in the education field. The system uses *neural network models* to extract concepts from raw data to build the KG.

For this work, given the complexity of collecting live weather and climate sensor data, there will be a primary focus on historical climate datasets for the North American region. Similarly,

TABLE I
DETAILS OF EACH DATASET (DATA SIZE IS MEASURED IN CSV FORMAT)

Data Source	Domain	Size
NOAA GHCND	Climate	5.1 GB
AviationStack	Flight	2.3 GB
Climateq	CO_2 emission	1.8 GB
Simplemaps	Administrative regions	4.5 GB

historical tourism datasets detailing tourist demand and movement, such as flight data, will also be used. Where applicable, Web APIs can be exploited as an option for collecting live and historical data on the following (we direct readers to this page[6] for an overview of sourced datasets): estimated climate emissions from the Climatiq API [31] and flight data from AviationStack [32]. In total, climate data are sourced from over 5000 weather stations across the United States and integrated with flight data from over 100 000 domestic flights to provide a knowledge base that displays the relationship between weather and flights. Further data are integrated to provide CO_2 emission estimations for flights to further demonstrate how climate-related data can be utilized in the modern context of climate change. An overview of the retrieved datasets is given in Table I.

2) Data Parsing and Cleaning: In order to collect weather data, a Python script made calls to the NOAA Web Services API [33], collecting weather data values in CSV format for each of the 5000 weather stations in the United States. NOAA provided a list of the unique identifiers of each weather station, which was used to retrieve the weather data on a station-by-station basis. An API key was not required to make a data request via URL and each request returned a CSV file that was appended to the end of the larger CSV dataset. The variables [5] collected included PRCP (Precipitation), SNWD (Snow Depth), SNOW (Snowfall), TAVG (Average Temperature), TMAX (Maximum Temperature), TMIN (Minimum Temperature), AWND (Average Wind Speed), WSF2 (Fastest 2-Minute Wind Speed). These values are stored as node properties of the *Weather* entity and were renamed to be more user friendly.

The AviationStack API [32] provides multiple endpoints for requesting data on airlines, airports, and flights. This was utilized to pull historical flight data for a specified date range in JSON format. The API response had to be parsed to filter out international flights that did not depart or arrive at a U.S. airport. The filtered dataset was then parsed from JSON to CSV format with redundant information removed to reduce the file size and improve parsing execution speed. Flight data included departure and arrival details such as takeoff time, airport codes, flight number, and airline carrier information.

Airport and airline data were also extracted using the AviationStack API to find all major U.S. airports and airlines. Each request returned 1000 results, and an offset was used to iterate over each value to check the departure and arrival airport International Civil Aviation Organization (ICAO) [34] code to verify if the result was a U.S. domestic flight. The ICAO and

---

[5]https://developers.booking.com/api/

[6]https://github.com/futaoo/climate-tourism-kg/blob/main/parsers/showtables.ipynb

```
    LOAD CSV WITH HEADERS FROM
↪   airports.csv AS line
            WHERE line.departure_airport
↪   IS NOT NULL
    MERGE (a:Airport {lat: line.latitude,
↪   long: line.longitude, iata:
            line.departure_iata, icao:
↪   line.departure_icao, name:
            line.departure_airport})
```

Listing 1.   Mapping airport entities

International Air Transport Association (IATA) [35] codes are standardized identifiers used in the aviation industry to track airports, flights and airlines and are stored as node properties for the *Airline* and *Flight* entities.

CO_2 flight emissions were estimated via the Climatiq API [31] to calculate the average CO_2 emissions released to transfer 100 passengers from the departure to arrival airport. Batch requests could be made using the "/batch" endpoint with up to ten values per request. First, the Google Maps GeoCoding API [36] was used to determine the latitude and longitude of corresponding U.S. airports. The geodesic distance in kilometers between two airports was then calculated using the Haversine formula [37] below:

$$d = 2r \sin^{-1}$$

$$\left(\sqrt{ \sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos \phi_1 \cos \phi_2 \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right) } \right) \tag{1}$$

where r is the radius of the Earth (6371 km), d is the distance between two points, ϕ_1 and ϕ_2 are the latitudes of the two points, and λ_1 and λ_2 are the longitudes of the two points, respectively. The calculated geodesic distance was passed to the Climatiq API as a parameter to generate a CO_2 estimation for each unique flight path. The type of estimation used was *passenger_flight-route_type_na-aircraft_type_na-distance*, which calculates a generic estimation for a domestic flight in the U.S. region for 100 passengers across the specified distance.

The cities and towns data were retrieved from a dataset [38] provided by SimpleMaps that contains over 28 000 cities and towns incorporated in the United States. In order to connect weather stations to nearby cities, a Python script retrieved the coordinates of each city with a population of over 10 000 people and built a K-dimensional tree of cities for searching. The weather station's latitude and longitude were then used to query the KD-Tree to find the closest city using the k-nearest neighbor algorithm. The same method described was used to connect airports to nearby cities. This produced intermediary CSV files for describing the relations between entities, particularly the "NEARBY" relation.

3) Ingesting the Graph: Once the appropriate CSV files were developed, the LOAD CSV command in CYPHER was used to ingest the datasets into the graph, describing the entities and relations step-by-step. Listing 1 illustrates how the *Airport* entity was created. Once ingested, the node property data types had

```
    MATCH (c:Airport)
    SET c.lat = toFloat(c.lat)
    SET c.long = toFloat(c.long)
```

Listing 2.   Mapping node properties

```
    LOAD CSV WITH HEADERS FROM
↪   city_station.csv AS line
            WHERE line.STATION IS NOT NULL
    MATCH (s:Station), (c:City)
    WHERE s.station = line.STATION AND
↪   line.CITY = c.name
    MERGE (s)<-[:NEARBY]->(c)
```

Listing 3.   Mapping the relationship between stations and cities

Fig. 3.   Relationship properties of a City to Weather Station.

to be set to ensure calculations on number data types, such as integers and floats, were efficient. Neo4j reads all CSV data as a string; thus, node properties were set using the following CYPHER command (see Listing 2).

In some cases, intermediate CSV files were created to describe relations between entities. Connecting *Station* to *City* involved an intermediary table to connect a *Station* ID to *City* name. Listing 3 demonstrates how this relation was ingested into the KG.

Following this, relationship properties, such as the calculated geodesic distance between a *Station* and *City* could be set. Fig. 3 demonstrates an example of one such relationship. This process was followed for each entity and relation in the KG with the relevant properties and data types set.

IV. EVALUATION AND RESULTS

The system developed for the climate–tourism KG presents a unique challenge in evaluating the success of this works. Knowledge graphs differ widely in scale, technology, and utilize various knowledge models. In order to evaluate this KG, a quantitative user survey was conducted using the Post-Study System Usability Questionnaire (PSSUQ), first proposed in [39] and developed by IBM in the 1990s. A total of 16 standardized questions were used to evaluate the following areas:

- Overall User Satisfaction.
- System Usefulness (SYSUSE).
- Information Quality (INFOQUAL).
- Interface Quality (INTERQUAL).

Results are calculated by averaging the results of the questions across each category. A score of 1 represents "Strongly Agree"

TABLE II
EVALUATION RESULTS ON PSSUQ METRICS

Overall User Satisfaction	1.85
System Usefulness (SYSUSE)	1.86
Information Quality (INFOQUAL)	1.89
Interface Quality (INTERQUAL)	1.81

and a score of 7 as "Strongly Disagree." A lower score indicates a higher level of user satisfaction.

A. Evaluation Scheme

The study was conducted using Google Forms [40] with access to a cloud-deployed sample of the KG on AuraDB [41]. Ten example queries were provided to the users to interact with the KG and familiarize themselves with the graph schema and CYPHER. Participants were selected as students from the School of Computer Science, University College Dublin, and had varying levels of knowledge in the domains of climate, knowledge graphs, and machine learning. Participants were contacted via word of mouth, and no personally identifiable information was collected.

Data collection occurred over a one-week period with participants opting to take part in a 10–15 min activity and survey. Users gained access to the KG via a private URL with an access username and password. From here, it was possible for participants to interact with the graph directly through CYPHER to execute the example commands. Participants were also able to create their own queries and manipulate the data in the KG. After each user completed the survey, the graph was reset to its original state to ensure a consistent experience for all users.

B. Results

The PSSUQ derives four calculated metrics to represent the overall user satisfaction of the developed knowledge graph. A total of 12 participants completed the activity and survey. The calculated scores are outlined in Table II.

The results showed consistency across all four categories of testing, with users reporting an overall satisfaction of 1.85. A score closer to 1 indicates participants had an extremely high level of satisfaction (Strongly Agree) with their user experience of the knowledge graph and the method of querying. A score closer to 7 indicates an extremely high level of dissatisfaction with the KG (Strongly Disagree). A score of 4 suggests neither satisfaction nor dissatisfaction (Neutral).

System usefulness (SYSUSE) scored similarly with a derived score of 1.86, bringing it in line with overall user satisfaction. This represents the participants' high level of satisfaction with the usefulness of the developed knowledge graph in terms of its practicality and the use cases outlined in the example queries. Participants were able to accurately modify the provided queries to derive new information from the knowledge graph in a practical and efficient manner.

The Information Quality (INFOQUAL) score was the lowest of the four scores at 1.89 by a small margin. Overall, the level of information provided to users allowed them to navigate the graph effectively. The user survey graph consisted of a sample
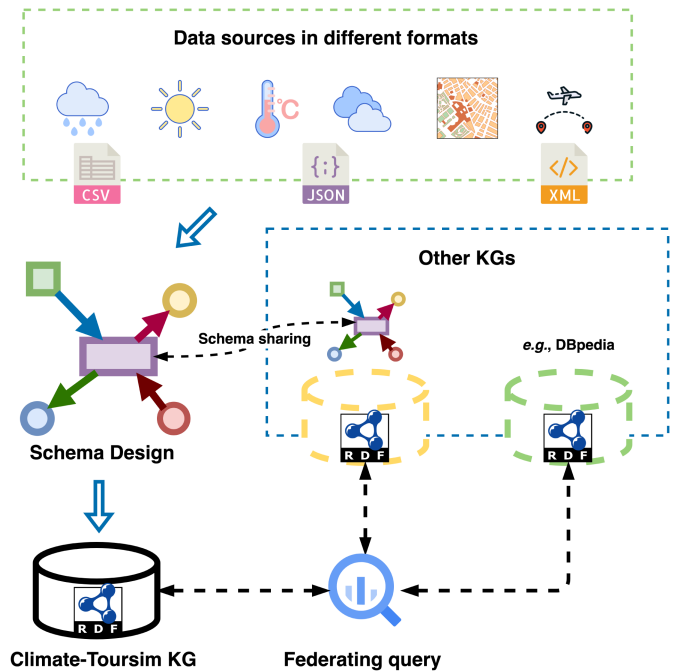

Fig. 4. Overview of data augmentation via knowledge graphs.

dataset for testing purposes, which may explain the lower score for INFOQUAL. The highest score was for Interface Quality (INTERQUAL), which scored 1.81. This reflects the quality of the Neo4j interface and the color coding and interaction applied to make the knowledge graph visually pleasing. Overall, there were high levels of reported user satisfaction across all four categories.

V. USE CASES OF THE CLIMATE–TOURISM KNOWLEDGE GRAPH

This section outlines some of the practical use cases of the KG explored for the purposes of this project. Specifically, this use case study aims to address some complicated tourism-related questions such as: Where should I go in January for a ski trip? What city should I visit to avoid busy airports? What is the most common flight paths for people visiting Los Angeles, and what are the estimated CO_2 emissions per passenger? A common challenge posed by these questions is that they can hardly be solved with individual data sources. To answer these questions, data consumers must invest considerable effort in augmenting the data. In contrast, once the aforementioned climate–tourism knowledge graph has been constructed, relevant data and features can be organized with a homogeneous schema. Data consumers can then use this single schema to extract features covering diverse data sources to answer these complex questions. Fig. 4 illustrates the need to convert different data formats to the knowledge graph (as shown by the hollow arrows) to achieve the augmentation and selection of homogeneous features. Furthermore, a profound method for augmenting remote data and features can also be accomplished when our climate–tourism schema is fully or partially adopted by other

```
    MATCH
↪   (a:Airline)-[:CARRIES]->(f)-[:EMITS]->
    (e:Emission)
    RETURN a.name, sum(e.totalCO2) as
↪   totalEmissions
    ORDER BY totalEmissions DESC LIMIT 5
```
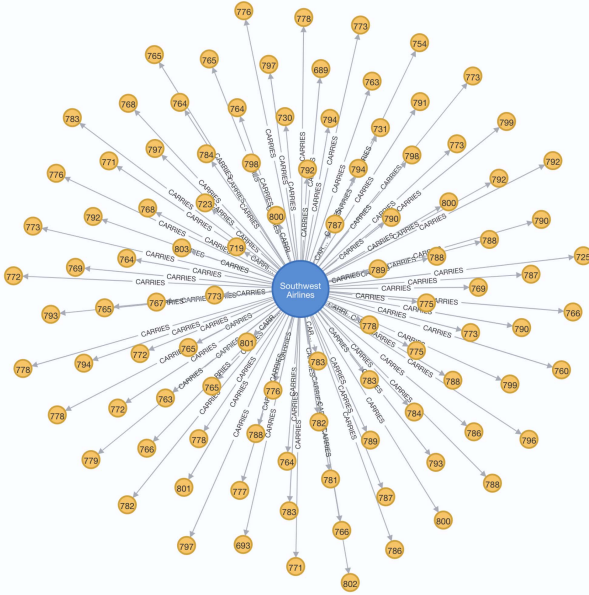
Listing 4.   Finding the airline entities with largest total CO_2 emissions

Fig. 5.   Case 1: Southwest Airlines and 100 flights.

TABLE III
TOTAL CO_2 EMISSIONS RANKED FOR DIFFERENT AIRLINES

Southwest Airlines	41 269 767 kg
American Airlines	31 234 857 kg
United Airlines	20 355 134 kg
Delta Air Lines	17 867 638 kg
Alaska Airline	14 234 668 kg

```
    MATCH
↪   (f:Flight)-[:EMITS]->(e:Emission)
    RETURN f.iata, max(e.totalCO2) as
↪   maxEmissions
    ORDER BY maxEmissions DESC LIMIT 1
```
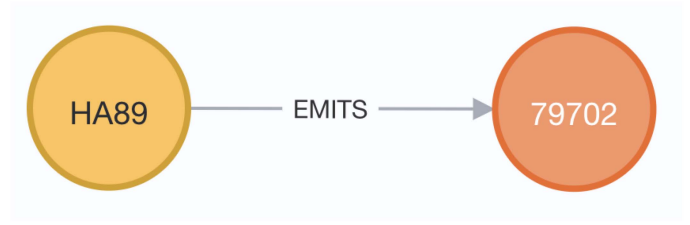
Listing 5.   Determining the flight with the largest CO_2 emissions

Fig. 6.   Query response.

```
    MATCH (w:Weather)
    WITH max(w.snowDepth) AS maxSnowDepth
    MATCH
↪   (c:City)-[:HAS_WEATHER]-(n:Weather
↪   {snowDepth: maxSnowDepth})
    RETURN c, n
```

Listing 6.   Largest single recorded value of snowDepth

KGs over the Semantic Web. In other words, multiple knowledge graphs published at distinct HTTP endpoints are inter-accessible (following the dashed arrows in Fig. 4). The following use case study will focus on the data and feature augmentation to answer complex questions from the perspective of only the proposed local climate–tourism KG.

A. Case 1: Airline CO_2 Emissions

The rise of climate-conscious consumers has led to increased transparency between businesses and their customers in regards to their CO_2-producing activities. Calculating CO_2 emissions for flights is widely used, such as in Google Flights [42]. Each *Flight* has a calculated CO_2 *Emission* associated with it through the EMITS relation. Listing 4 exploits this relationship to find the *Airline* entities with the largest total CO_2 emissions in the KG using the *sum* aggregate function. For example, Fig. 5 shows the CO_2 emissions of 100 flights to be aggregated into Southwest Airlines. Table III shows a descending list of airlines by total CO_2 emissions, with Southwest Airlines accounting for the majority of CO_2 pollution.

Consumers could use the KG to better inform themselves about how airline companies contribute to domestic CO_2 emissions within the United States. Airlines could utilize these data to improve flight route planning and become more environmentally conscious when planning new routes. The *Flight* with the largest CO_2 emissions is Flight HA89 from Boston to Honolulu, the longest domestic flight in the US, emitting 79 702 kg of CO_2, according to Listing 5 and its response in Fig. 6.

B. Case 2: Let us Go Skiing

The prospective question "Where should I go for a ski trip in January?" was proposed as a potential use case and an aim for a developed KG in the climate–tourism domain. In planning a trip that is reliant on weather conditions, whether it be rainfall for a camping trip or levels of solar radiance for a beach day, KGs can offer insight to make better-informed, data-driven decisions. Listing 6 searches the graph for the largest single recorded value of snowDepth in the KG. The HAS_WEATHER (see Fig. 7) relationship is exploited to connect the weather entity to a nearby place.

Fig. 7. Query response.

```
    MATCH
→   (c)-[:ARRIVES]-(a:Airport)-[:DEPARTS]-(f)
    RETURN a, COLLECT(a) as
→   leastBusyAirports
    ORDER BY SIZE(leastBusyAirports) ASC
→   LIMIT 3
```

Listing 7. Finding the least busy airports

Fig. 8. Query response.

C. Case 3: Busy Airports

Many airports are centrally located, acting as travel hubs for international flights. Well-known examples of these airports include Dubai International Airport outside of the United States and Hartsfield-Jackson Atlanta International Airport (ATL) in the United States. Finding travel routes through the least busy airports indicates the areas of lesser popularity in terms of air travel. Listing 7 queries the least busy airports from the KG, and the results for the top three least busy airports are shown in Fig. 8. This type of query could benefit a consumer who is searching for quieter destinations that avoid metropolitan areas. Airlines could use this information to determine their least popular airports when deciding to change and remove routes.

```
    CALL
→   gds.pageRank.stream(airportsAndFlights)
    YIELD nodeId, score
    RETURN gds.util.asNode(nodeId).name AS
→   name, score
    ORDER BY score DESC, name ASC
```

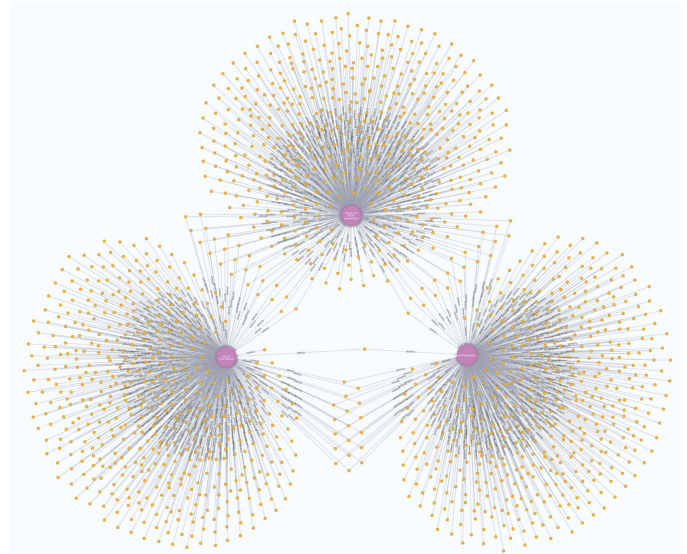Listing 8. Sorting airports by the number of departing flights

Fig. 9. Case 3: Top three airports with most departing flights.

Kenai Airport is located in the remote town of Kenai, Alaska. Watertown International Airport is located in Jefferson County, NY, USA. It is used for general aviation and commercial flights are subsidized by the U.S. government through the Essential Air Service Program [43]. Lastly, Chrisholm/Range Regional Airport is located in Hibbing, MN, USA, with a town population of around 16 000 people. Similarly, it is possible to find the *Airports* with the most departing flights in a given period. This query (Listing 8) can be executed using the Page Rank algorithm using the Graph Data Science Library in Neo4j [44]. This query returns a sorted list of *Airports* by the calculated Page Rank score.

ATL had the highest Page Rank score at 90.93. Denver International Airport and Dallas/Fort Worth International Airport scored 85.95 and 76.45, respectively. The three busiest airports in the United States according to the number of departing flights are depicted in Fig. 9. The visualization also reveals the shared flights between these *Airports*.

D. Case 4: Route Distance

The Euclidean distance between entities such as *Station* and *City* is presented in the relationship properties of the KG. Listing 9 finds the *Weather* entity with the highest recorded maximum temperature on 1 January 2022, and returns the distance value between the *Station* and *City*.

```
    MATCH (w:Weather {date: "2022-01-01"})
    WITH max(w.maxTemp) AS
↪   highestTemperature
    MATCH (c:City)-[a:NEARBY]-(s:Station)-
↪   [:OBSERVES]-(z:Weather{maxTemp:
↪   highestTemperature})
    RETURN z.station, z.date,
↪   highestTemperature, a.distance, c.name
```

Listing 9. Retrieving the *Weather* entity with the highest recorded maximum temperature and returning the distance to the *City*
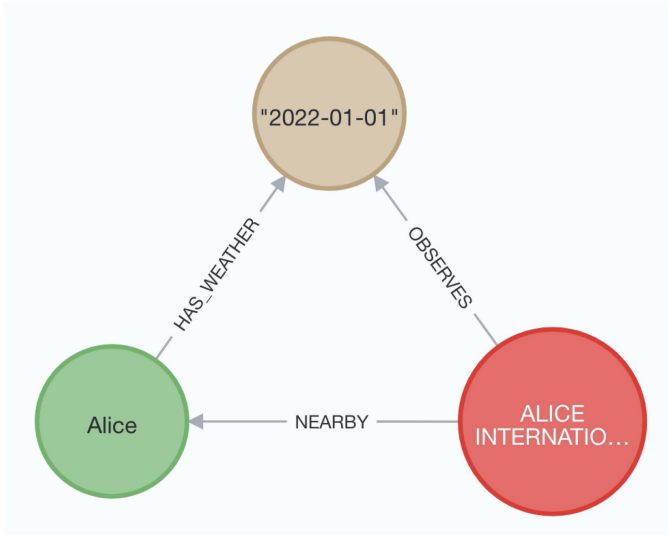

Fig. 10. Case 4: Query response.

```
    MATCH (a:City)-[b:NEARBY]-(c)-[d:DEPAR
↪   TS]-(e:Flight {iata:
↪   AC3049})-[f:ARRIVES]-(g)-[h:NEARBY]-(i
↪   :City)
    RETURN b.distance + d.distance +
↪   h.distance
```

Listing 10. Distance calculation from a sub-graph path

Fig. 11. Subgraph path.

This returns the calculated distance of 2.95 km between Alice, TX, USA, and the weather station located at Alice International Airport (Fig. 10). The inclusion of these distances can allow for more complex operations, such as calculating the total distance of a subgraph or path of entities. For example, Listing 10 calculating the distance between the city of Alice, Alice International Airport, Dallas/Fort Wort International Airport, and the city of Dallas (see Fig. 11). This use case has many practical applications; from a consumer perspective, this KG could be used to calculate travel distances between entities in the KG, such as between a *City*, *Airport*, and an *Airport* and *City* somewhere else in the United States where the two *Airports* are connected by a *Flight*.

VI. CURRENT LIMITATIONS

Although we received some positive feedback regarding the system's usability and the use case study demonstrates the system's usefulness for relevant applications in the area, we believe there is still considerable room for improvement in the future with regards to the system's effectiveness of the climate–tourism KG. In the following sections, we highlight the key considerations that scholars who are interested in the same issues as us should take into account.

A. Data Limitations

Weather and flight data are both data-rich resources with widespread availability. There are over 100 000 daily flights moving up to 12 000 000 people around the globe everyday according to FlightRadar24 [45]. Similarly, in the United States, the Cooperative Observer Network has access to volunteered weather stations numbering over 8500 [46]. These stations comprise the "climate network," which records measures of rainfall, snow, wind, sunlight, and temperature. These instantaneous measures made at predetermined times are often processed further for one-day observations, which are subsequently forwarded to the National Climatic Data Center for processing and archiving (e.g., publishing data as daily summaries). Processing data at this scale is resource-intensive in terms of processing power and storage. This presents the main limitation of this work. Further limitations include the frequency of climate data reporting. Daily weather summaries for the U.S. geographic region only become widely available in the days after observation. Hourly summaries are provided at a similar rate. In order to derive real-time insights for active weather events and flights, the frequency of this reporting would have to be improved to allow for live data to be ingested into the KG.

B. Historical and Real-Time Data

Working with real-time data presents significant challenges when dealing with climate data. Datasets are often provided in hourly or daily summaries, and inconsistencies in the type of sensors used in weather stations lead to missing values and irregularities. Similarly for flight data, there is a wealth of real-time flight information available today, largely due to the development of a variety of flight APIs, such as AviationStack [32], FlightRadar24 [45], and FlightAware [47]. Contrary to weather data, detailed and consistent historical flight data prior to the previous decade are difficult to source. Flight data are generally useful for real-time applications such as flight tracking [47], and historical use cases tend to be less common. In order to build a comprehensive KG in this domain, historical flight data from multiple sources would need to be standardized and processed to ensure consistency across the KG.

C. Storage Requirements

The storage capacity requirements are a limiting factor in the development of KGs at scale. The quantity of weather and flight data generated daily is significantly prior to being integrated into a connected KG. A comparison between the sample dataset that was utilized for the quantitative user survey shows that storing the integrated data in a graph database uses 260% more storage than storing the data in CSV tabular format. The survey KG was a total of *approx.* 10.8 MB, compared with 4.1 MB for the same data in plain CSV. While the storage increase is highly significant, it offers increased practicality by integrating and relating data in a semantic manner, more closely representative of the real world, and in a way that is otherwise not queryable in an efficient manner; thus, there is a storage:practicality tradeoff to be considered when developing KGs.

VII. CONCLUSION

The proposed climate–tourism knowledge graph was developed in Python, CYPHER and Neo4j. Working with live data at scale presents an opportunity to improve the efficiency of data collection and cleaning, either by utilizing a lower level language such as Java as part of the data collection process or by capitalizing on the Neo4j API and program drivers to interact with the KG remotely and automate the graph ingestion process. The current knowledge graph architecture acts as a proof of concept and practicality and should act as a starting point for a real-world graph application in this field. Our findings include the development of an effective method of parsing Weather, Flight, and Emissions data from NOAA [33], AviationStack [32], and Climatiq [31] and the conversion of these datasets into a graph-based DBMS. The usability study, conducted using the PSSUQ [39] approach, demonstrates the practicality and usefulness of KGs as a solution to connect previously unrelated data in an integrated system that is queryable, simple to use, and efficient.

In the future, we plan to adapt the architecture to incorporate live data sources, which would provide a real-time application for this field that is yet to be developed and utilized. Working with real-time data does pose specific challenges, particularly due to the frequency of weather reporting. Incorporating a live weather source, such as OpenWeatherMap.org [48], would broaden the scope for future applications of the KG. Furthermore, we will concentrate more on the application of semantic web techniques in recommender systems. To achieve this, the climate–tourism knowledge graph will be developed to interlink with open knowledge graphs (e.g., Wikidata, DBPedia) as well as some existing travel data providers (e.g., TripAdvisor) where user-related data, such as reviews, are available. We will then explore to what extent the power of Linked Data can enhance the performance of existing tourist recommender systems. On the other hand, this iteration of the KG focuses on the North American geographic region due to the availability of large weather datasets from NOAA [33] and to allocate the limited research resources for a specific country, acting as a proof of concept for developing knowledge graphs in this domain. Expanding the geographic scope to a continental or international level would require incorporating a much larger number of data

sources into a standardized format. We would explore additional resources in terms of storage and processing power to progress to a KG of this scale.

REFERENCES

[1] H. T. Oo, W. W. Zin, and C. C. T. Kyi, "Assessment of future climate change projections using multiple global climate models," *Civil Eng. J.*, vol. 5, no. 10, pp. 2152–2166, 2019.

[2] M. Lynas, B. Z. Houlton, and S. Perry, "Greater than 99% consensus on human caused climate change in the peer-reviewed scientific literature," *Environ. Res. Lett.*, vol. 16, no. 11, 2021, Art. no. 114005.

[3] A. Bigano, J. M. Hamilton, and R. S. Tol, "The impact of climate change on domestic and international tourism: A simulation study," *SSRN Electron. J.*, vol. 7, Jul. 2006, doi: 10.2139/ssrn.907454.

[4] D. Scott, S. Gössling, and C. M. Hall, "International tourism and climate change," *Wiley Interdiscipl. Rev.: Climate Change*, vol. 3, no. 3, pp. 213–232, 2012.

[5] M. S. Pathan, J. Wu, Y. H. Lee, J. Yan, and S. Dev, "Analyzing the impact of meteorological parameters on rainfall prediction," in *Proc. IEEE USNC-URSI Radio Sci. Meeting (Joint AP-S Symp.)*, 2021, pp. 100–101.

[6] H. Li, H. Song, and L. Li, "A dynamic panel data analysis of climate and tourism demand: Additional evidence," *J. Travel Res.*, vol. 56, no. 2, pp. 158–171, 2017.

[7] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[8] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "An ontology model for climatic data analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 5739–5742.

[9] M. Chein and M.-L. Mugnier, *Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs.* Berlin, Germany: Springer, 2008.

[10] J. Wu, F. Orlandi, D. O'Sullivan, E. Pisoni, and S. Dev, "Boosting climate analysis with semantically uplifted knowledge graphs," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4708–4718, 2022.

[11] J. Wu, H. Chen, F. Orlandi, Y. H. Lee, D. O'Sullivan, and S. Dev, "Automated climate analyses using knowledge graph," in *Proc. IEEE USNC-URSI Radio Sci. Meeting (Joint With AP-S Symp.)*, 2021, pp. 106–107.

[12] C.-I. Lee, T.-C. Hsia, H.-C. Hsu, and J.-Y. Lin, "Ontology-based tourism recommendation system," in *Proc. 4th Int. Conf. Ind. Eng. Appl.*, 2017, pp. 376–379.

[13] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts.* Pennsylvania, PA, USA: IGI Global, 2011, pp. 205–227.

[14] R. Novak et al., "Harmonization and visualization of data from a transnational multi-sensor personal exposure campaign," *Int. J. Environ. Res. Public Health*, vol. 18, no. 21, 2021, Art. no. 11614.

[15] G. Weikum, "Knowledge graphs 2021: A data odyssey," *Proc. VLDB Endowment*, vol. 14, no. 12, pp. 3233–3238, 2021.

[16] J. L. Medina-Franco, G. M. Maggiora, M. A. Giulianotti, C. Pinilla, and R. A. Houghten, "A similarity-based data-fusion approach to the visual characterization and comparison of compound databases," *Chem. Biol. Drug Des.*, vol. 70, no. 5, pp. 393–412, 2007.

[17] J. Wu, F. Orlandi, Y. H. Lee, D. O'Sullivan, and S. Dev, "Organizing decentralized energy data using semantic approach," in *Proc. Photon. Electromagn. Res. Symp.*, 2021, pp. 2213–2216.

[18] "SEAS ontology," Accessed: Jun. 10, 2022. [Online]. Available: https://ci.mines-stetienne.fr/seas/index.html

[19] J. Wu, F. Orlandi, D. O'Sullivan, and S. Dev, "Detecting rainfall events leveraging climate knowledge graphs," in *Proc. Photon. Electromagnetics Res. Symp.*, 2021, pp. 2336–2341.

[20] J. Wu, F. Orlandi, T. AlSkaif, D. O'Sullivan, and S. Dev, "Ontology modeling for decentralized household energy systems," in *Proc. Int. Conf. Smart Energy Syst. Technol.*, 2021, pp. 1–6.

[21] J. Wu, H. Chen, F. Orlandi, Y. H. Lee, D. O'Sullivan, and S. Dev, "An interoperable open data portal for climate analysis," in *Proc. IEEE USNC-URSI Radio Sci. Meeting (Joint With AP-S Symp.)*, 2021, pp. 104–105.

[22] "Google knowledge graph search API," Accessed: Jul. 11, 2022. [Online]. Available: https://developers.google.com/knowledge-graph

[23] R. Wang et al., "AceKG: A large-scale knowledge graph for academic data mining," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 1487–1490.

[24] "Diffbot," Accessed: Jul. 11, 2022. [Online]. Available: https://www.diffbot.com/

[25] D. Sullivan, "A reintroduction to our knowledge graph and knowledge panels," May 2020, Accessed: Jul. 11, 2022. [Online]. Available: https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/

[26] J. O. Costa and A. Kulkarni, "Leveraging knowledge graph for open-domain question answering," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2018, pp. 389–394.

[27] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data: The story so far," in *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. Pennsylvania, PA, USA: IGI Global, 2011, pp. 205–227.

[28] D. Allemang, F. Gandon, and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL*. New York, NY, USA: ACM Books, 2020.

[29] "Khan Academy," Accessed: Jul. 11, 2022. [Online]. Available: https://www.khanacademy.org/

[30] P. Chen, Y. Lu, V. W. Zheng, X. Chen, and B. Yang, "Knowedu: A system to construct knowledge graph for education," *IEEE Access*, vol. 6, pp. 31553–31563, 2018.

[31] "Climatiq - carbon emissions calculation API," Accessed: Jul. 11, 2022. [Online]. Available: https://www.climatiq.io/

[32] "aviationstack—Real-Time flight tracker & status API," Accessed: Jul. 11, 2022. [Online]. Available: https://aviationstack.com/

[33] "NCEI Web Services," Accessed: Feb. 11, 2022. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/webservices

[34] "ICAO," Accessed: Feb. 27, 2022. [Online]. Available: https://www.icao.int/Pages/default.aspx

[35] "IATA," Accessed: Jul. 11, 2022. [Online]. Available: https://www.iata.org/

[36] "Google maps platform documentation," Accessed: Jul. 11, 2022. [Online]. Available: https://developers.google.com/maps/documentation/geocoding

[37] N. C. Soe and T. L. L. Thein, "Haversine formula and RPA algorithm for navigation system," *Int. J. Data Sci. Anal.*, vol. 6, no. 1, 2020, Art. no. 32.

[38] "US cities database," Accessed: Jul. 11, 2022. [Online]. Available: https://simplemaps.com/data/us-cities

[39] J. R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *Int. J. Human-Comput. Interact.*, vol. 7, no. 1, pp. 57–78, 1995.

[40] "Google forms: Online form creator," Accessed: Jul. 11, 2022. [Online]. Available: https://www.google.com/forms/about/

[41] D. Fitzpatrick, "Neo4j aura – managed cloud graph database service," Accessed: Jul. 11, 2022. [Online]. Available: https://neo4j.com/cloud/platform/aura-graph-database/

[42] "Google flights," Accessed: Mar. 15, 2022. [Online]. Available: https://www.google.com/travel/flights

[43] "Essential air service," Accessed: Jul. 11, 2022. [Online]. Available: https://www.transportation.gov/policy/aviation-policy/small-community-rural-air-service/essential-air-service

[44] "The Neo4j graph data science library manual v2.1," Accessed: Jul. 11, 2022. [Online]. Available: https://neo4j.com/docs/graph-data-science/current/

[45] Flightradar, "Flightradar24—Real-time flight tracking," May 2021, Accessed: Jul. 11, 2022. [Online]. Available: https://www.flightradar24.com

[46] "Cooperative observer network (COOP)," Accessed: Nov. 15, 2022. [Online]. Available: https://www.ncei.noaa.gov/products/land-based-station/cooperative-observer-network

[47] "FlightAware—Flight tracker / flight status," Accessed: Jul. 11, 2022. [Online]. Available: https://uk.flightaware.com/

[48] "API Weather," Accessed: Jul. 11, 2022. [Online]. Available: https://openweathermap.org/api
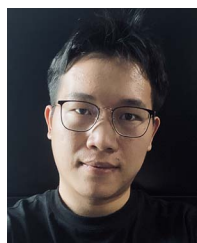
Jiantao Wu received the M.Sc. degree in materials for energy and environment from University College London, London, U.K., in 2017.

He is currently a Ph.D. Student under the joint supervision of Prof. Soumyabrata Dev at University College Dublin, Dublin, Ireland, and Dr. Fabrizio Orlandi and Prof. Declan O'Sullivan at Trinity College Dublin, Dublin. He was a Software Engineer for two years with China Electronics Technology Group Corporation until 2019. His research interests include knowledge graphs, machine learning, and sensor data processing.

Jarrett Pierse received the B.Sc. degree in computer science from University College Dublin, Dublin, Ireland, in 2022.

He is currently a Systems Development Engineer with Data Centre Networks, Amazon, Dublin, Ireland. His research interests include business and technology, particularly how we can leverage technology to improve the everyday lives of people and tackle some of the biggest challenges facing humanity today.

Fabrizio Orlandi received the B.Eng. and M.Eng. degrees in computer engineering from Universitá degli Studi di Modena e Reggio Emilia, Modena, Italy, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from the University of Galway, Galway, Ireland, in 2013.

He is currently a Marie Skłodowska-Curie EDGE Fellow with Trinity College Dublin, Dublin, Ireland. He currently leads the project "DynamoKG"—exploring dynamic and uncertain facts in knowledge graphs. He is also involved in the "Beyond 2022" project as Senior Researcher working on KGs for Irish History. Previously, since 2018 at ADAPT, he has also led the research agenda of two industry projects with Wolters Kluwer and Huawei. Prior joining ADAPT—at Fraunhofer IAIS in Germany—he had the role of Coordinator in the EU H2020 OpenBudgets.eu project and contributed to large European and industry research projects, such as BigDataOcean.eu and SLIPO.eu. He has experience on foundational and applied research on both EU-funded and industry projects. His research interests include knowledge graphs, linked open data, knowledge representation, and the application of semantic technologies to different domains, such as social media, cultural heritage, law, and open data.

Declan O'Sullivan received the B.A. (Mod), M.Sc., and Ph.D. degrees in computer science from Trinity College Dublin (TCD), Dublin, Ireland, in 1985, 1988, and 2006, respectively.

He is currently a Professor in Computer Science with the School of Computer Science and Statistics, Trinity College Dublin (TCD), Dublin, Ireland, and is a co-applicant Principal Investigator with the ADAPT SFI Research Centre. Since joining TCD from industry in 2001, he has established himself as an international research leader in his field. He has authored more than 260 scientific peer-reviewed papers and international journals.

Prof. O'Sullivan is a member of three journal editorial boards, and was the Chair for more than 12 IEEE and IFIP conferences over the years. He has won competitive research funding as PI and Co-PI of approximately 7.8M Euros. He has won funding across a range of funding programs: European Commission (H2020 and Marie Curie); Science Foundation Ireland (FAME, CNGL, ADAPT); HEA PRTLI (NEMBES, TGI), and from industry: Huawei, Accenture, Ericsson, Nokia Bell Labs, Ordnance Survey Ireland, Central Statistics Office. He was elected as a Fellow in Trinity College Dublin in 2019 in recognition for the quality of his contributions.

Soumyabrata Dev (Member, IEEE) received the B.Tech. degree (*summa cum laude*) in electronics and communication engineering from the National Institute of Technology Silchar, Silchar, India, in 2010, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2017.

He is currently an Assistant Professor with the School of Computer Science, University College Dublin, Dublin, Ireland, and an SFI Funded Investigator with ADAPT SFI Research Centre, Dublin. In 2015, he was a Visiting Student with Audiovisual Communication Laboratory (LCAV), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. He has authored more than 120 publications in leading journals and conferences. His research interests include remote sensing, statistical image processing, machine learning, and deep learning.