Data & Knowledge Engineering 156 (2025) 102405

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Automated mapping between SDG indicators and open data: An LLM-augmented knowledge graph approach

Wissal Benjira [a,b,*], Faten Atigui [c,**], Bénédicte Bucher [b,**], Malika Grim-Yefsah [b,**], Nicolas Travers [a,c,**]

[a] Léonard de Vinci Pôle Universitaire, Research Center, 92 916 Paris La Défense, France
[b] Conservatoire National des Arts et Métiers, CEDRIC-CNAM, France
[c] LASTIG, Université Gustave Eiffel, IGN, Paris, France

ARTICLE INFO

Keywords:
Sustainable Development Goals (SDG)
Large language model (LLM)
Knowledge graph (KG)
Open data
Schema mapping

ABSTRACT

Meeting the Sustainable Development Goals (SDGs) presents a large-scale challenge for all countries. SDGs established by the United Nations provide a comprehensive framework for addressing global issues. To monitor progress towards these goals, we need to develop key performance indicators and integrate and analyze heterogeneous datasets. The definition of these indicators requires the use of existing data and metadata. However, the diversity of data sources and formats raises major issues in terms of structuring and integration. Despite the abundance of open data and metadata, its exploitation remains limited, leaving untapped potential for guiding urban policies towards sustainability. Thus, this paper introduces a novel approach for SDG indicator computation, leveraging the capabilities of Large Language Models (LLMs) and Knowledge Graphs (KGs). We propose a method that combines rule-based filtering with LLM-powered schema mapping to establish semantic correspondences between diverse data sources and SDG indicators, including disaggregation. Our approach integrates these mappings into a KG, which enables indicator computation by querying graph's topology. We evaluate our method through a case study focusing on the SDG Indicator 11.7.1 about accessibility of public open spaces. Our experimental results show significant improvements in accuracy, precision, recall, and F1-score compared to traditional schema mapping techniques.

1. Introduction

The Sustainable Development Goals (SDGs), established by the United Nations (UN), serve as a global call to action to address contemporary challenges and promote a sustainable future [1]. The 17 goals cover a wide range of social, economic, and environmental aspects. Each goals is accompanied by specific targets and indicators designed to monitor progress [2]. Measuring these indicators relies heavily on access to relevant information from open datasets scaled in both time and space. The UN Statistical Commission considers only a limited number of these indicators to be widely available [3].

One of the main challenges related to the assessment of SDGs progress lies in the vast amount of data generated from various sources. Confronted to this complexity, international organizations are urging nations to enhance documentation and data collection efforts [1]. Thus, heterogeneity, semantic consistency and interoperability of data are a long way from being granted [4].

* Corresponding author at: Léonard de Vinci Pôle Universitaire, Research Center, 92 916 Paris La Défense, France.
** Corresponding authors.
E-mail addresses: wissal.benjira@devinci.fr (W. Benjira), faten.atigui@lecnam.net (F. Atigui), benedicte.bucher@ign.fr (B. Bucher), malika.grim-yefsah@ensg.eu (M. Grim-Yefsah), nicolas.travers@devinci.fr (N. Travers).

https://doi.org/10.1016/j.datak.2024.102405
Received 14 June 2024; Received in revised form 12 November 2024; Accepted 30 December 2024
Available online 3 January 2025

Knowledge Graphs (KGs) emerge as a promising solution to overcome the hurdles associated with heterogeneous data sources and formats [5]. Indeed, by representing data as interconnected nodes and relationships, KGs provide a flexible and scalable framework for integrating diverse datasets. KGs have been applied to SDGs in projects such as `LinkedSDGs` [6] and `SustainGraph` [4]. These projects mainly focus on goals and targets interactions, yet indicator assessment remains largely unexplored.

To compute SDG indicators, a UN documentation corpus details each formula with related metadata. This metadata is provided as unstructured textual documents and gives a clear description of indicator computation requirements. However, manually processing and analyzing these documents can be time-consuming and is prone to errors. There is need to automate the extraction and analysis of information from these documents. Additionally, linking these documents with sources demands deep domain expertise. Therefore, it is guiding us towards using Natural Language Processing (NLP) techniques such as entity recognition and textual mapping.

Recently, following the advancements in NLP, mechanisms with enhanced understanding of language semantics have been able to achieve better results. Large Language Models (LLMs) have made significant strides in tackling complex tasks that demand a profound understanding of semantics. Recent researches have explored integrating LLMs to augment KGs to consider the textual information [7] but none of them has been applied in the domain of sustainability. It represents a promising tool to associate disparate data to the measurement of an SDG Indicator. Leveraging LLMs for schema mapping has been shown to provide initial promising results, principally due to their ability to understand and interpret the semantics and context of the data schemas.

To improve LLM performance, we suggest using advanced filtering rules during pre-processing. These rules include regular expressions for pattern matching, type checks to ensure data consistency, and value range validation to filter out anomalies. By implementing these steps, the data is cleaned and structured more effectively before being fed into the LLM, significantly enhancing the quality and relevance of the results produced. This approach ensures that the LLM works with more accurate and reliable data, leading to better overall performance in various tasks. Going further, we propose to apply collection rules to ensure spatial indicator and conceptual disaggregation.

This paper presents two main contributions. First, we propose a new approach to define an SDG schema for indicators' measurement based on indicators metadata and open data in order to structure the KG from both open data source SDG point of views. We implement the result into a graph in order to manage the relations between concepts represented as nodes and to allow a global view for the mapping process as well as the indicator computation. Second, we introduce a method for automatically mapping data sources to SDG indicators using LLMs and pre-filtering rules. The obtained associations are added to the graph and form our resulting KG.

The rest of the paper is structured as follows. The next section reviews related work. Section 3 presents our global approach and modeling method. Section 4 describes the conceptualization step and Section 5 presents the LLM schema mapping one. Section 6 describes the experimental evaluation, presents the results, and discusses the insights and limitations obtained. Finally, Section 7 concludes the paper and open with perspectives.

2. Literature review

This section examines the literature on SDG data modeling, Natural Language Processing techniques, and the potential of Large Language Models and Knowledge Graphs to enhance the analysis and integration of open data.

2.1. Challenges and opportunities in SDG data modeling

The SDGs offer an analytical grid in which it is relevant to project decision-making. Various research activities have attempted to develop and apply techniques capable of managing the complexities associated with SDG data [8–11]. However, very few studies have proposed computational models to represent their links with external open data. Efforts to achieve the SDGs rely on modeling and simulation, creating model-based decision support systems [12–14]. The authors in [15] have introduced a framework based on graph theory to model the SDGs, offering a systemic view of their interdependencies. KGs have been applied to Sustainable Development Goals (SDGs) in projects such as LinkedSDGs [6] and SustainGraph [4], which mainly focus on goals and targets interactions. However, the area of indicator assessment remains largely unexplored.

In France, the National Institute of Statistics and Economic Studies (INSEE) coordinates the collection of statistical indicators to track progress. The French National Council for Statistical Information (CNIS) has set up a working group dedicated to the SDGs, highlighting priority targets lacking indicators. Recommendations, in particular for SDG 11 *"Sustainable Cities and Communities"* guide towards priority indicators. A research work is suggested around accessibility indicators. We can estimate it by taking into account urban characteristics that define or influence traffic conditions, in particular: distance from public transport, employability and diversity of land use [16]. These three factors are identified by INSEE's open data sources. However, that have to get structured in order to answer to one or more SDG indicator.

2.2. Large Language Models (LLMs)

Since indicators and related formulas are described in long documentations, it advocates the usage of Natural Language Processing (NLP) techniques to extract main concepts. Moreover, the link between those concepts and the corresponding data for its computation implies some quality measurements.

LLMs have shown great performance in various NLP tasks. Indeed, these models, capable of understanding and generating human-like text, are trained on vast amounts of data. Their ability to understand context and semantics has made a significant advancements in how machines process and interpret languages.

Generative LLMs, such as GPT, are designed to generate human-like text and can be fine-tuned for specific tasks or reused across various applications without additional training [17–20]. Embedding models, including BERT and Ada [21], offer contextual embeddings that have significantly enhanced performance across numerous NLP tasks [22,23]. Moreover, using embeddings for text representation facilitates efficient and accurate passage retrieval through semantic similarity [21,24].

However, LLMs trained on general corpora often struggle to generalize well to specific domains or new knowledge due to the lack of domain-specific knowledge or new training data [25]. To address this, pre-processing with filtering rules such as regex, type checks, and value range validation can help clean and structure the data before using LLMs, enhancing their performance and relevance for specialized tasks [26].

Despite their success in many applications, LLMs have been criticized for their lack of factual knowledge. Specifically, LLMs memorize facts and knowledge contained in the training corpus [27].

2.3. LLMs for Knowledge Graphs (KGs)

Recent research has explored integrating LLMs to augment KGs to consider the textual information and improve the performance in downstream tasks [7]. KGs emerge as a promising solution to overcome the hurdles associated with heterogeneous data sources and formats [5]. By representing data as interconnected nodes and relationships, KGs provide a flexible and scalable framework for integrating diverse datasets.

Researchers take advantage of LLMs to process the textual corpus in the KGs and then use the representations of the text to enrich KGs representation [28]. Some studies also use LLMs to process the original corpus and extract relations and entities for KG construction [29].

Recent studies try to design a KG prompt that can effectively convert structural KGs into a format that can be comprehended by LLMs. In this way, LLMs can be directly applied to KG-related tasks such as KG construction [30], KG completion [31] and KG reasoning [32]. KG completion refers to the task of inferring missing facts in a given knowledge graph. KG construction involves creating a structured representation of knowledge within a specific domain. This includes identifying entities and their relationships with each other. KG reasoning, on the other hand, involves deriving new insights and knowledge from existing data by applying logical inference over the relationships and entities within the KG. All these recent technique have not yet been applied to the studied domain because of the lack of structure in both open data and indicator documentation.

2.4. Schema mapping

Schema mapping is a fundamental task in data management and integration, involving the identification of semantic correspondences between elements of two or more database schemas [33]. However, the schema mapping task is challenging due to several inherent complications. Firstly, schemas are designed with different perspectives and terminologies, namely textual heterogeneity, reflecting the conceptualization of domain experts from disparate fields. This semantic heterogeneity can lead to ambiguous mappings where schema elements have the same name but different meanings, or different names but the same meaning. Secondly, structural heterogeneity compounds this complexity, with schemas exhibiting varied architectures, hierarchies, constraints, and model granularity differences.

Recently, following the advancements in NLP, mechanisms with enhanced understanding of language semantics have been able to achieve better results. Utilizing LLMs for schema mapping has shown promising initial results, mainly because of their ability to understand and interpret the semantics and context of data schemas [34,35].

A major problem of many automatic schema mapping approaches is that they fail if the semantic similarity is hard to detect [26]. For example, instance-based column mapper typically fail to map columns that contain disjoint but semantically similar values such two tables with different street names or even worse the same content in different languages (*e.g.,* French and English) [36]. More recent approaches also try to consider semantic aspects such as synonyms and hypernyms or rely on Machine Learning (ML) [26].

Hättasch et al. (2020) have shown that neural word embeddings can be utilized to propose a small set of possible candidates for schema matching which is crucial for data integration. Instance-based approaches work well with different types of entities. Weaknesses are found, for example, with attributes that all contain human names: the embeddings are good for finding other attributes with names, but a further subdivision is difficult. We propose to enhance the schema mapping process by adding mappings based on disaggregation criteria. Disaggregation involves breaking down attributes into finer components, which can be crucial for accurately mapping complex data structures. For example, an attribute representing the population could be disaggregated into components such as sex, age, and region. By incorporating disaggregation rules, we ensure that the LLM can identify and align not only the primary attributes but also their constituent parts. This approach leverages the semantic understanding of LLMs to detect relationships that might be missed by traditional existing methods.

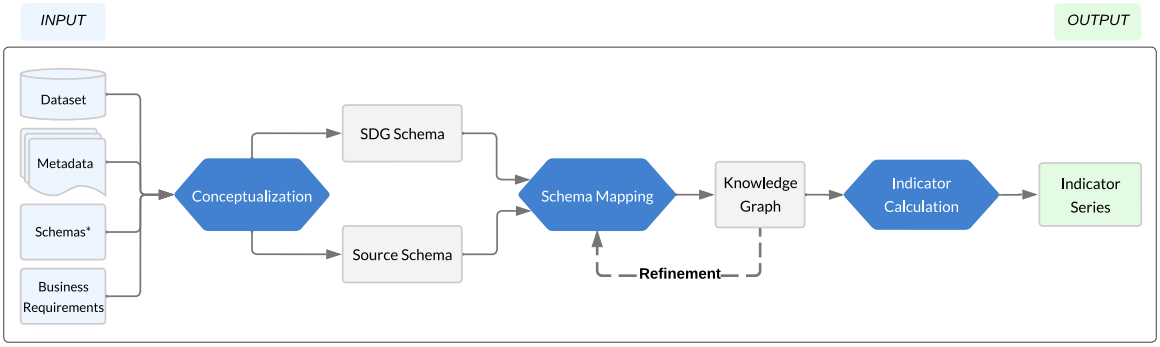

Fig. 1. Approach overview for SDG indicator modeling and calculating using open data.

2.5. Synthesis and positioning

None of the current methods fully address the generic computation of indicators from open data. LLMs are powerful but struggle with domain-specific tasks due to their training on general corpora. Pre-processing with filtering rules can enhance their performance. KGs provide a structured way to represent data, highlighting relationships between entities. This aligns with our goal of mapping connections between data sources and the SDG indicators. Integrating LLMs with KGs can enrich the graph with relationships from data sources and aid in schema mapping. This combined approach leverages the strengths of both LLMs and KGs for effective indicator computation using open data.

3. Approach overview

In this section, we present an overview of our approach for SDG indicator modeling and calculating using Open Data (Fig. 1). This approach is divided into three main phases: Conceptualization, Schema Mapping, and Indicator Computation. The process begins with the collection of the inputs, including mainly dataset from diverse sources such as government databases and international organizations, the metadata, providing information about one or more aspects of the data. Besides, as inputs, we can also have business requirements and commonly used schema. These inputs are essential as they form the foundation for the entire data modeling effort. The selection of data sources is driven by specific the business requirements, which outline the conceptualization objectives and needs.

The `Conceptualization` process focuses on generating data schema from the heterogeneous inputs. This step is crucial as it sets the stage for how the data will be organized and integrated in subsequent steps. During this phase, data elements are defined, and their interrelationships are established to form a general data model. The schemas generated in this phase provide a detailed representation of data structures, including entities, attributes, and their relationships. On the one hand, the SDG metadata descriptions provided by United Nation documentation are employed to identify the main entities used in the indicator formula, resulting in a defined list of entities and their attributes. These new concepts contribute to the SDG indicator model, enriching the global SDG schema. On the other hand, data sources are defined and structured, typically open data from governmental institutions. The output of the conceptualization phase is a set of detailed schemas, including the SDG schema and the data sources schema necessary to compute the indicator. The specific methodologies and technical aspects of schema generation are explained in Section 4.

In the `Schema Mapping` phase, we align the SDG Schema with existing Data Sources Schema. The goal is to identify semantic correspondences between schema elements. This task is challenging due to textual and semantic heterogeneity. We employ LLMs combined with pre-filtering rules to enhance performance. LLMs help in identifying potential matches by interpreting the context and meaning of schema elements. Pre-filtering rules, such as regex, type checks, and value range validations, clean and structure the data before processing by LLMs. These techniques reduce noise and improve accuracy. In this part, we also include rules to allow disaggregated attributes. The output of this phase is a set of new relationships and associations between the SDG schema and the data sources schema. These connections enrich the Knowledge Graph (Section 5).

In the `Indicator Calculation` phase, we compute SDG indicators using the link mapped in the KG. We start by extracting relevant data from the enriched knowledge graph. Using these connections, we apply predefined formulas to calculate the SDG indicators. This process helps assess progress towards the SDGs with integrated and structured data and produces the needed indicator series. The quality of this computation is presented in Section 6.

4. Conceptualization

Modeling Open Data with a view to integrating it into a management or analysis system requires a systematic and methodical approach. We have developed a three-stage approach to guide this process (Fig. 2). We present the followed approach in Section 4.1 and two use case application in Section 4.2.

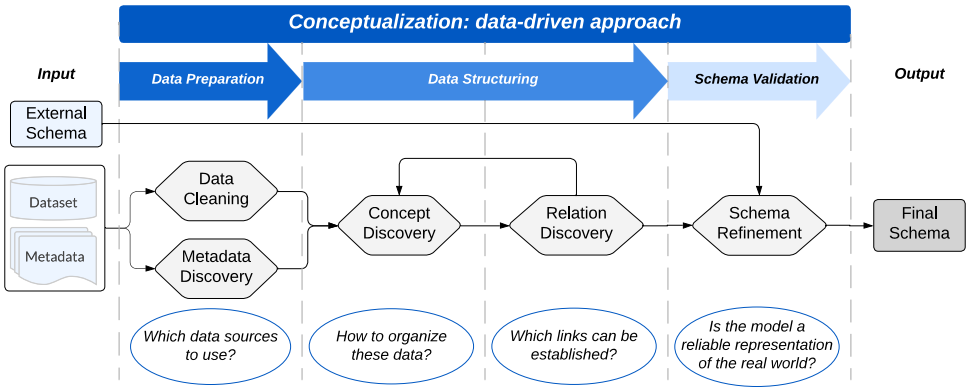

Fig. 2. Data-driven approach for modeling open data.

4.1. Data-driven modeling approach

This section outlines a systematic and manual, three-stage data-driven approach for modeling data (cf Fig. 2).

`Data preparation` is the initial phase referrers to data pre-processing where data is prepared from various sources. The business requirements specified by domain experts and stakeholders guide the choice of data and metadata sources relevant to the research [37]. Then, the two inputs are processed in this phase, which includes Data Cleaning, which involves fixing errors, removing duplicates, and handling missing values. Another important step is Metadata Discovery, where metadata is identified to provide context about the data. For SDG indicators, we use SDG-specific metadata, while for other datasets, metadata describes the attributes of each data source. This stage prepares the ground by gathering the raw materials needed to build a solid data model [38].

Once the data has been collected, the next step is `Data Structuring` those data. Concept Discovery identify the key entities or elements in the data [39]. This includes obtaining data fields, grouping relevant entities, detecting constraints and identifying gaps in data quality. This phase is crucial for organizing the raw data into meaningful entities and identifying the essential features to be integrated into the model. Then, the schema's structure is created by establishing relationships, aggregations, association structuring and compositional links between identified entities. Relation Discovery aims to discover new relations from a given text corpus without annotated data [40]. This phase requires a thorough understanding of the data and its context to define how it interacts with each other. It aims to capture the dependencies and interactions between entities to best reflect reality.

Finally, `Schema Validation` is a step that ensures that the data structure meets predefined standards. This includes both syntactic validation, such as ensuring data type consistency, and semantic validation, which confirms that the data accurately represents real-world entities and relationships. This process includes schema refinement, where the data schema is improved based on external insights. It also involves integrating relevant external data sources in the external schema step to enhance the analysis. Finally, the process culminates in the final schema, where the refined and validated data schema is completed. This ensures data integrity and consistency, making the data schema reliable for real-world representation [41].

Following these steps, we aim to create open data models that are both accurate and tailored to specific needs. These steps are essential to ensure the integration of open data into the management or analysis system. This manual methodology enables us to convert raw data into usable information.

4.2. Constructing conceptual models

In this section, we apply the approach proposed for the construction of a conceptual schema for SDG and for a use case about sport event in urban environments. Both approaches illustrate the importance of abstractly representing essential elements and relationships within various domains.

4.2.1. Conceptualization for SDG

Data Preparation: To build the first model, the data and metadata collected come from INSEE, French government data for monitoring the SDGs, and open data and documentation provided by the United Nations. A manual study of SDG indicator metadata repository has been done. It reflects the latest reference metadata information provided by the UN System and other international organizations on data and statistics around SDG indicators.

Data Structuring: The key concepts retained are *impact*, *challenge*, *goal*, *target*, *indicator* and *series*. According to the Agenda 2030, the *goals* have a *impact* multidimensional, touching at the same time environmental, economic and social aspects. They cover one or more of the challenges defined by the Agenda 2030 roadmap. A *goal* expresses an ambitious but specific commitment, and always begins with a verb or an action. Each *goal* is made up of a number of *target*, which have results that can be quantified by an *indicator*, which is a precise measure that makes it possible to assess whether a *goal* has been achieved. Each *goal* is enriched by

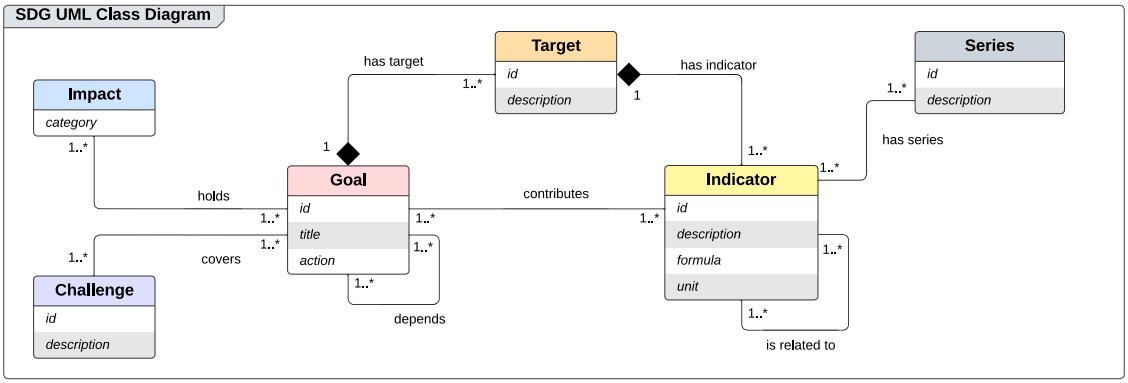

Fig. 3. SDG UML class diagram.

indicators. In rare cases, the same indicator may belong to several goals. Finally, a *series* is a set of observations on a quantitative characteristic that provides concrete measures for an indicator. Each *series* contains several records of data points organized by time, geographical area or other dimensions of interest (gender, age range, etc.).

Schema Validation: The schema (Fig. 3) has been completed with correspondences using notions from the ontologies *Linked SDG* [9] and *Sustain Graph* [4]. The conceptual schema (Fig. 3) represents the elements of the domain and their relationships, making explicit the relevant domain knowledge.

To illustrate the instantiation of the conceptual model, we use Goal 11 on sustainable cities as an example. With half of the world's population living in urban areas, urban challenges are central to this goal. Structuring open data for the SDGs, particularly in urban management and environmental sustainability, is essential to addressing these complex challenges. Cities are rich in open data, which provides a valuable resource to guide urban policies towards more sustainable and resilient pathways. Open data modeling has therefore become a crucial component of this transformation.

Accordingly, Fig. 4 presents an excerpt of the graph (508 nodes in total) for Goal 11 untitled *"Sustainable cities and communities"*. This goal aims to make cities and human settlements inclusive, safe, resilient, and sustainable. This goal includes specific Targets 11.7, which focuses on providing safe access to green spaces and public areas, emphasizing universal access for vulnerable groups such as women, children, older persons, and persons with disabilities. The Indicator 11.7.1 object is linked to this target and measures the average share of the built-up area of cities that is open space for public use, categorized by sex, age, and disability status. The formula for this indicator is the ratio of the total population within 400 meters of service areas to the total population within the city, expressed as a percentage. The UN metadata defines concepts and their description to compute the indicator. For computing the formula of this indicator the three following concepts are needed: *City*, *PublicSpace* and *Street*. The challenges covered by Goal 11 are given by the 2022 French Volunteer National Revue and as described in a UN report, the impacts are categorized into *Economic*, *Social*, and *Environmental* types.

In order to instantiate the output schema, we store data as a Labeled Property Graph, a type of Knowledge Graph (KG). KGs have gained popularity in recent years for their ability to organize large, complex datasets into easily navigable and semantically rich structures. We find KGs ideal for representing interconnections among SDGs, as they structure entities, relationships, and attributes in a graph format. This graph data structure helps to handle both models origin, and especially mappings between the SDG Entities. The motivation of this choice is twofold. First, the graph structure helps the exploration between indicators relationships. Second, the graph serves as a valuable tool for refining indicators. Analysis of the graph's topology enables the identification of opportunities for refinement, while mapping errors are easily identifiable.

4.2.2. Conceptualization for a use case

Among the opportunities supported by cities, urban sport event stand out as emblematic use case. Indeed, the events, whether national or international sporting championships, create a complex environment with imperatives of logistics, safety, sustainability and social impact. Sporting events can act as a catalyst for sustainability projects.

The modeling of sports infrastructures and host sites plays a crucial role in event planning. Grim-Yefsah et al. (2019–2020) have proposed a conceptual schema to represent the characteristics of an sport event and the links between it and sports infrastructures. Wu et al. (2017) and Kurowska et al. (2021) have used geographic information systems to map these infrastructures, taking into account factors such as capacity, location and accessibility. According to Preuss, there is currently a gap in the literature when it comes to modeling and integrating sport event concepts into information systems or databases.

Therefore, we propose to tackle this issue by applying our method (Fig. 2) to build a conceptual schema for structuring of sport event data.

Data Preparation: The data and metadata collected come from multiple sources listed in Table 1. Mainly governmental data sources:

- `Datagouv`: it is the open, community-based platform designed to centralize and structure open data in France.

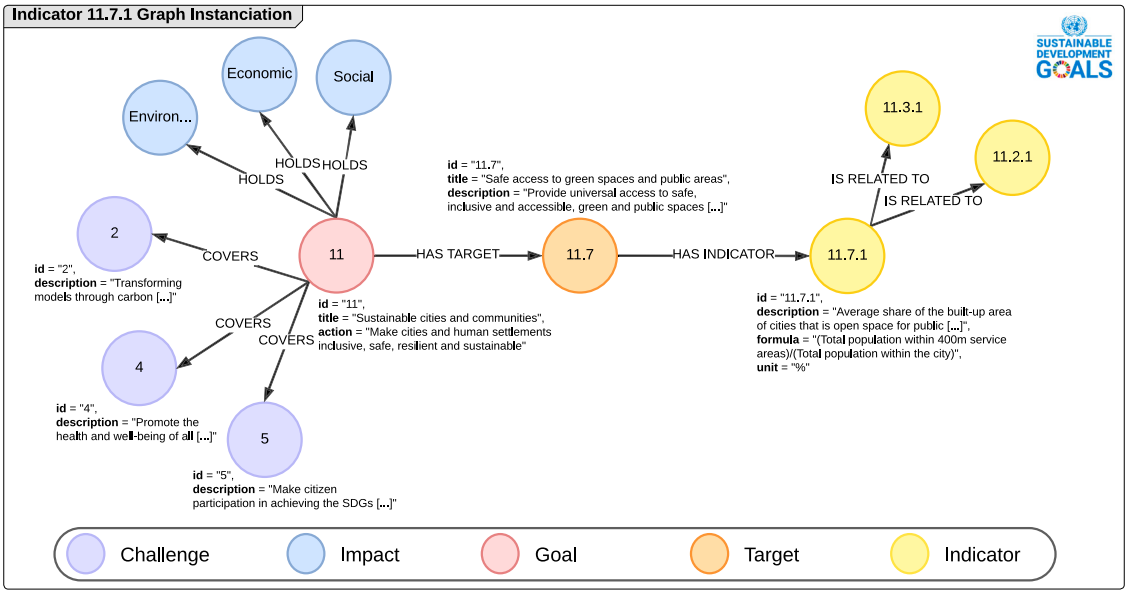

Fig. 4. Indicator 11.7.1 graph instantiation.

Table 1
Sources and descriptions of collected data.

Source	Description	Year
Datagouv	Census of sports facilities, practice areas, and sites	2021
IDF Mobility	Stations of the Île-de-France region rail network	2023
IGN	Geographic referential of the territory concerning the most recent IRIS zones	2023
IGN	Linear description of the Île-de-France IGN railway network	2023
INJEP	Data from sports federations	2022
INSEE	Data from the Base permanente des équipements en France	2021
French Ministry of Sports	Database of sports equipment families and types	2023

- IDF Mobility: Île-de-France Mobilités is the organizing authority for sustainable mobility in Île-de-France region.
- IGN: The National Institute of Geographic and Forest Information is a public administrative establishment placed under the joint authority of the Ministries in charge of ecology and forestry. It is the reference public operator for geographic and forest information in France.
- INJEP: The National Institute for Youth and Popular Education is a national agency of the French Ministry of Education and Youth. INJEP is both a knowledge-producing observatory and a center for resources and expertise on youth issues and policies, popular education, associative life and sport.
- INSEE The National Institute of Statistics and Economic Studies collects, analyzes and disseminates information on the French economy and society.

Data Structuring: The main concepts retained are *federation*, *sporting event*, *sporting equipment*, *facility*, *address* and *transportation*. The schema integrates a multitude of entities, relationships and attributes, capturing different aspects of SE, from geographical information to details of sports facilities, sports federations and participant data. The schema highlights the interconnectivity of these elements, underscoring their impact on the planning, logistics and management of SE.

Schema Validation: The schema (Fig. 5) was completed with the public amenities data schema from the public data schema repository. The conceptual schema presented in Fig. 5 offers a holistic view of sport event data.

The conceptual schema developed, illustrated in Fig. 5, was applied to the *2016 European Football Championship* case study. The schema was converted into a data graph by means of "direct" mapping. Stored in *Neo4j*, this graph (Fig. 6) was explored using *Cypher* queries, providing a global view of interconnected entities.

This graph (Fig. 6) serves as an illustrated example of the sport event class diagram (Fig. 5) and illustrates the relationships between different entities associated with *2016 European Football Championship*. The event was organized by the *French Football Federation* and uses several sport equipments belonging to different sport installations *e.g.*, *Stade Vélodrome*, and *Stade de France*. These stadiums and fields are crucial venues for the *2016 European Football Championship* matches and other related events. Each stadium and field is located in specific *IRIS Zone* (represented the fundamental unit for dissemination of infra-municipal data). In this *IRIS Zone*, are located various public transport stops, for instance *Saint Denis-Porte de Paris* and *Basilique de Saint Denis*.

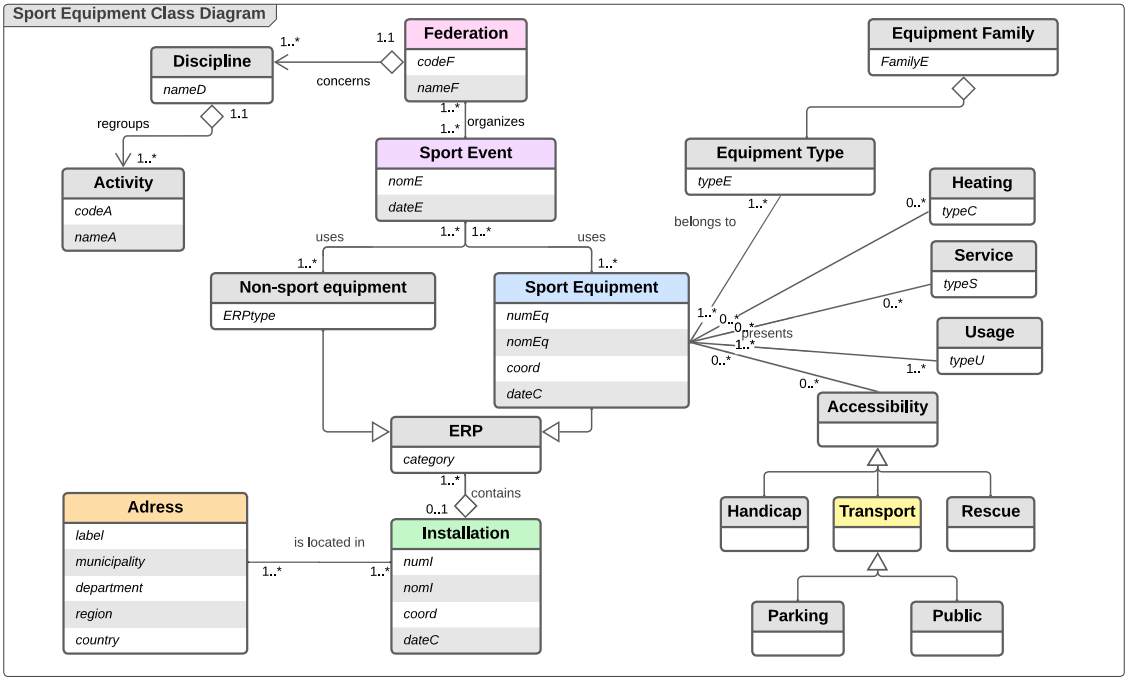

Fig. 5. Sport event class diagram.

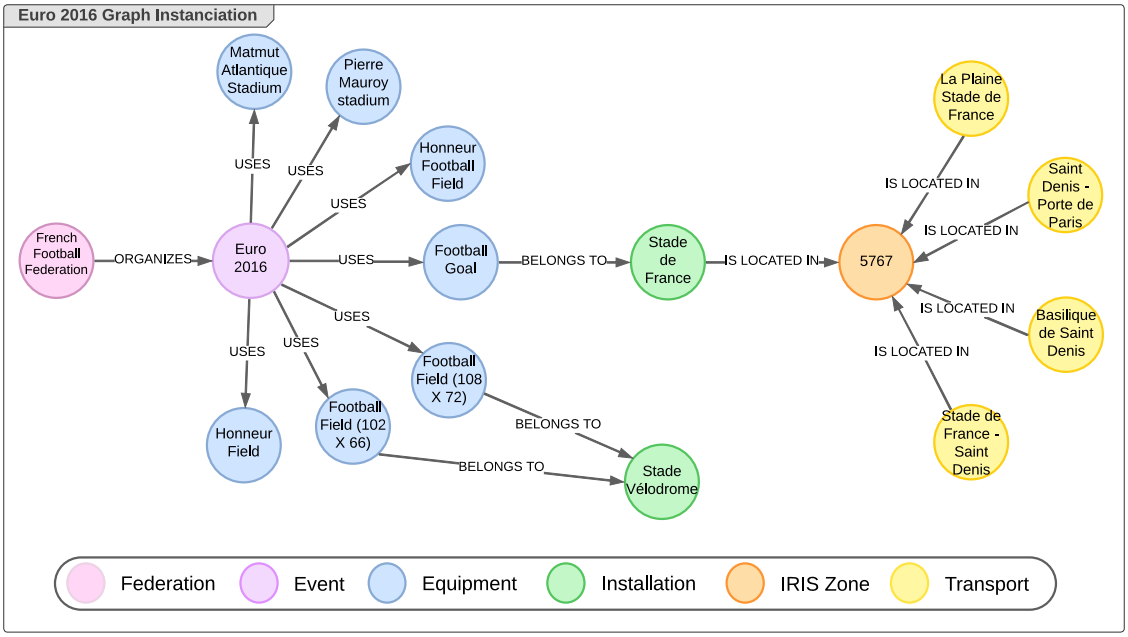

Fig. 6. Graph instantiation for Euro 2016 event.

4.2.3. Observation

As shown in Figs. 4 and 6, the two schemas we obtained are currently disassociated, meaning they are not yet linked or integrated. Our objective is to calculate an indicator from open data, which requires connecting these schemas. With the structuring framework now established, our next step is to identify and establish correspondences between the schemas. This integration process will enable us to use the data effectively and proceed with calculating the desired indicator accurately.

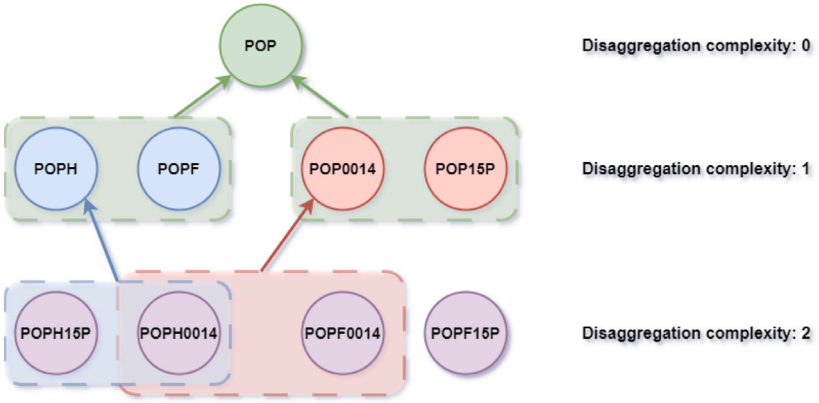

Fig. 7. Disaggregation example on population attributes.

5. Schema mapping

In this section, we introduce our approach using the advances of LLMs to tackle the issue of schema mapping. Indeed, leveraging LLMs for schema mapping offers significant advantages thanks to their ability to understand and interpret the semantics and context of data. LLMs can automatically identify and align corresponding elements across schemas. To further enhance and achieve more accurate schema mapping, we propose to defining filtering rules to guide the LLM, in order to improve its precision and efficiency in identifying relevant schema elements. A key specification of our work is the focus on attribute disaggregation, allowing a data structure at a more granular level to support detailed analysis.

We propose a novel approach for schema mapping based on LLM and represented as a Knowledge Graph. The main idea is to use a two-step approach consisting of a rule-based filtering step followed by an attribute matching step. In both steps we use embeddings on different levels. Thus, in we propose an algorithm which integrates these predefined rules and LLM-based comparisons to create an enriched knowledge graph. In Section 5.1, we define the attribute disaggregation process, which enables a more granular mapping of schema elements. The Section 5.2 introduces the refining rules to guide the LLM in schema alignment. Finally, in Section 5.3, we detail the full algorithm, which integrates disaggregation, refining rules, and LLM-based comparisons for enhanced schema mapping.

5.1. Disaggregation

The global SDG indicator framework has an overarching principle of data disaggregation: "*Sustainable Development Goal indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics*". To achieve this, there is a need to monitor and assess SDG indicators based on their disaggregation. Such a challenge is also highlighted in the SDGs geospatial roadmap provided by the UN Statistics Division [1].

Definition 1 (*Disaggregation*)**.** The breakdown of observations within a common branch of a hierarchy to a more detailed level to that at which detailed observations are taken. With standard hierarchical classifications categories can be split (disaggregated) when finer details are required and made possible by the codes given to the primary observations.

Definition 2 (*Disaggregation Dimensions*)**.** The characteristics by which data is to be disaggregated (by geographical location, sex, age, disability, etc.).

Definition 3 (*Disaggregation Categories*)**.** The different characteristics under a certain disaggregation dimension (female/male, etc.).

The Fig. 7 illustrates an example of disaggregation for the Population attribute. At the top level, we have the total population (POP). This is disaggregated into male (POPH) and female (POPF) subcategories, which are further broken down into age groups. For instance, the male population is divided into those aged 0–14 (POPH0014) and those aged 15 and above (POPH15P). Similar disaggregations apply to the female population (POPF0014 and POPF15P).

This hierarchical structure allows to check whether the sum of the subcategories matches the higher-level totals, thus ensuring the correctness and completeness of our data. The complexity of disaggregation increases with each level, from simple binary splits (complexity 1) to more detailed breakdowns (complexity 2).

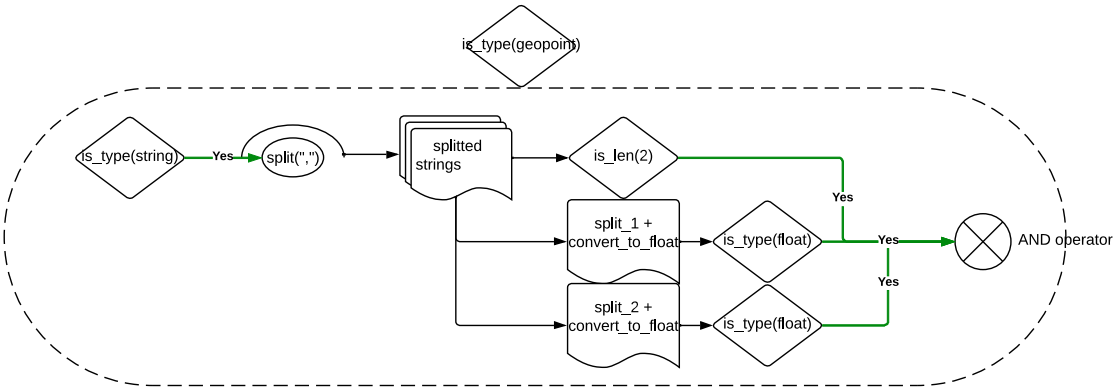

Fig. 8. Mapping rule example for *geopoint* type.

5.2. Rule-based filtering

To guide and improve the accuracy of the LLM mapping results, we propose a control process prior to LLM processing, known as Rule-Based Filtering [26]. These rules guide the process of aligning entities and attributes between the two schemas.

Definition 4 (*Rule-Based Filtering*)**.** Pre-processing method that systematically applies specific control rules to input data before it is processed by the LLM. This process includes:

- Type Consistency: Ensuring that mapped entities and attributes are of compatible types. This involves converting data types where necessary (e.g., converting strings to floats) to maintain consistency across mappings.
- Regular Expressions: Using patterns to validate and filter potential matches by identifying attributes that follow specific formats, such as date or geo-point formats.
- Value Range Limits: Excluding unlikely matches by comparing numerical attributes based on their expected ranges, thus eliminating values that fall outside of feasible limits.
- Averages and Statistical Checks: Where applicable, comparing values to averages or expected norms as an additional filtering measure.

By integrating these filtering steps, we assume that the accuracy of the LLM mapping results can be enhanced. Only semantically and syntactically relevant matches are considered, leading to a more reliable schema alignment.

For example, to verify that an attribute is of type *geopoint* (cf. Fig. 8), the rule begins by checking that the attribute is of type string. This ensures that the value can be processed as text. The next step involves splitting the string by a comma, which should result in exactly two parts. This is necessary because a *geopoint* typically consists of a latitude and a longitude, separated by a comma. After the split operation, the rule verifies that the result contains exactly two elements, confirming the expected format for a *geopoint*. Finally, each of these elements is checked to ensure they are of type float, which validates that both latitude and longitude are in the correct numerical format. By combining these checks, the rule effectively ensures that the attribute is correctly identified as a *geopoint*.

5.3. Algorithm description

The algorithm is designed to find semantic correspondences between elements of two or more database schemas based on LLM generation, by producing prompts containing source schema attributes with target schema candidates. This approach integrates predefined filtering rules and updates the knowledge graph with the new mapping relationships. The main steps of our method are as follows: (1) extracting schema descriptions in the graph, (2) applying predefined mapping rules, (3) leveraging LLMs for transforming the graph into a knowledge graph, and (4) filtering the results based on disaggregation criteria. For each function, we have provided a table detailing its parameters and description, enabling a clear understanding of each component's role in the algorithm.

The process begins by initializing a Knowledge Graph ($\mathcal{K}_G$ - Line 1) that includes the source schema (S_1) on the one hand, and the target schema (S_2) on the other hand. The descriptions for both schemas are extracted and represented as properties in the graph with each attribute's textual description and type (Lines 2–3). For each attribute in the source schema (Lines 4–24), the algorithm applies filtering rules to select potential candidates in the target schema based on defined criteria such as the type. These predefined matches form the initial set of potential matches (C_M).

Next (Lines 7–11), the algorithm leverages the LLM to compare the source attribute ($\mathcal{A}_S$) with each potential target attribute ($\mathcal{A}_T$), computing a similarity score ($\mathcal{M}_{score}$) for each pair. If a score exceeds the threshold τ, it is considered a valid match. The

algorithm keeps track of the top k potential matches for each source attribute. Each of these matches is then added to the graph as a new link, with the mapping score as a weight (Lines 13–15).

If an attribute in the source schema can be disaggregated (Lines 16–23), the algorithm performs an additional LLM-based check to see if the target attribute is a possible disaggregation of the source attribute. If this check passes, the disaggregation relation is also added to the graph. The final result is an enriched Knowledge Graph ($\mathcal{K}_G$) that includes all the new mapping relations identified by the LLM and the predefined rules. This graph not only contains the initial schemas but is enriched by the mappings and their associated scores.

6. Evaluation

This section presents the evaluation of our approach, detailing the experimental setup, driving example, and the methods used to assess the performance and accuracy of schema mapping using LLMs and traditional methods.

6.1. Experimental setup and driving example

For our evaluation, we consider the subgraph which is a graph instantiation of the SDG schema presented in Section 4. Fig. 9 shows how the attributes Position and Population are identified based on their definitions and rules. Position is identified as a geopoint, while Population is identified as a float with additional disaggregation rules (e.g., sex, age, disability). These attributes and their rules are predefined and validated in advance by SDG experts for each indicator.

In the following, we particularly focus on the SDG Indicator 11.7.1 related to the accessibility to open public spaces. For that, we used two primary datasets . The first dataset is derived from the French population census by INSEE since 2006. We kept only the columns that remained consistent from 2006 to 2020. Our study focused on the Île-de-France region and included various levels of aggregation. The second dataset comes from the French Sport Ministry and provides an open database related to sport facilities.

Next, we use `Pydantic` for its compatibility with `LangChain` in information extraction and classification. `Pydantic` acts as an output parser that defines a structured schema for LLM responses. By specifying a model in `Pydantic`, we ensure that outputs from the LLM are formatted according to our schema, which includes predefined types and validation rules. This approach helps maintain consistency and reliability in the extracted information, as each output conforms to a structured format. Detailed descriptions within the schema also support clearer and more accurate LLM interpretations.

Our approach involves rule-based filtering for mapping and disaggregation of mapped attributes. For each (entity, data source) pair, we check which attributes of the entity are eligible for mapping against each attributes of the data source. If any attributes

Algorithm 1 Schema Mapping Using LLM

Require: Source schema S_1, Target schema S_2, LLM $\mathcal{F}$, Threshold τ
Ensure: Knowledge Graph $\mathcal{K}_G$

1: $\mathcal{K}_G \leftarrow \text{InitializeGraph}(S_1, S_2)$
2: $\mathcal{D}_S \leftarrow \text{ExtractDescriptions}(S_1)$
3: $\mathcal{D}_T \leftarrow \text{ExtractDescriptions}(S_2)$
4: **for** $\mathcal{A}_S \in S_1$ **do**
5: $\quad C_M \leftarrow \text{ApplyFilteringRules}(\mathcal{A}_S)$
6: $\quad \mathcal{M} \leftarrow \{\}$
7: $\quad$ **for** $\mathcal{A}_T \in C_M$ **do**
8: $\quad\quad \mathcal{M}_{score} \leftarrow \text{Compare}(\mathcal{F}, \mathcal{A}_S, \mathcal{A}_T)$
9: $\quad\quad$ **if** $\mathcal{M}_{score} > \tau$ **then**
10: $\quad\quad\quad \mathcal{M}_i \leftarrow \text{TopKMatches}(\mathcal{M}_i, \mathcal{A}_T, \mathcal{M}_{score})$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: $\quad$ **for** $\mathcal{M}_i \in \mathcal{M}$ **do**
14: $\quad\quad \mathcal{K}_G \leftarrow \text{EnrichGraph}(\mathcal{K}_G, \{(\mathcal{A}_S, \mathcal{M}_i, \mathcal{M}_{score})\})$
15: $\quad$ **end for**
16: $\quad$ **if** $\text{IsDisaggregable}(\mathcal{A}_S)$ **then**
17: $\quad\quad$ **for** $\mathcal{M}_i \in \mathcal{M}$ **do**
18: $\quad\quad\quad \mathcal{M}_{score} \leftarrow \text{CheckDisaggregation}(\mathcal{F}, \mathcal{A}_S, \mathcal{M}_i)$
19: $\quad\quad\quad$ **if** $\mathcal{M}_{score} > \tau$ **then**
20: $\quad\quad\quad\quad \mathcal{K}_G \leftarrow \text{EnrichGraphWithDisaggregation}(\mathcal{K}_G, \{(\mathcal{A}_S, \mathcal{M}_i, \mathcal{M}_{score})\})$
21: $\quad\quad\quad$ **end if**
22: $\quad\quad$ **end for**
23: $\quad$ **end if**
24: **end for**
25: **return** $\mathcal{K}_G$

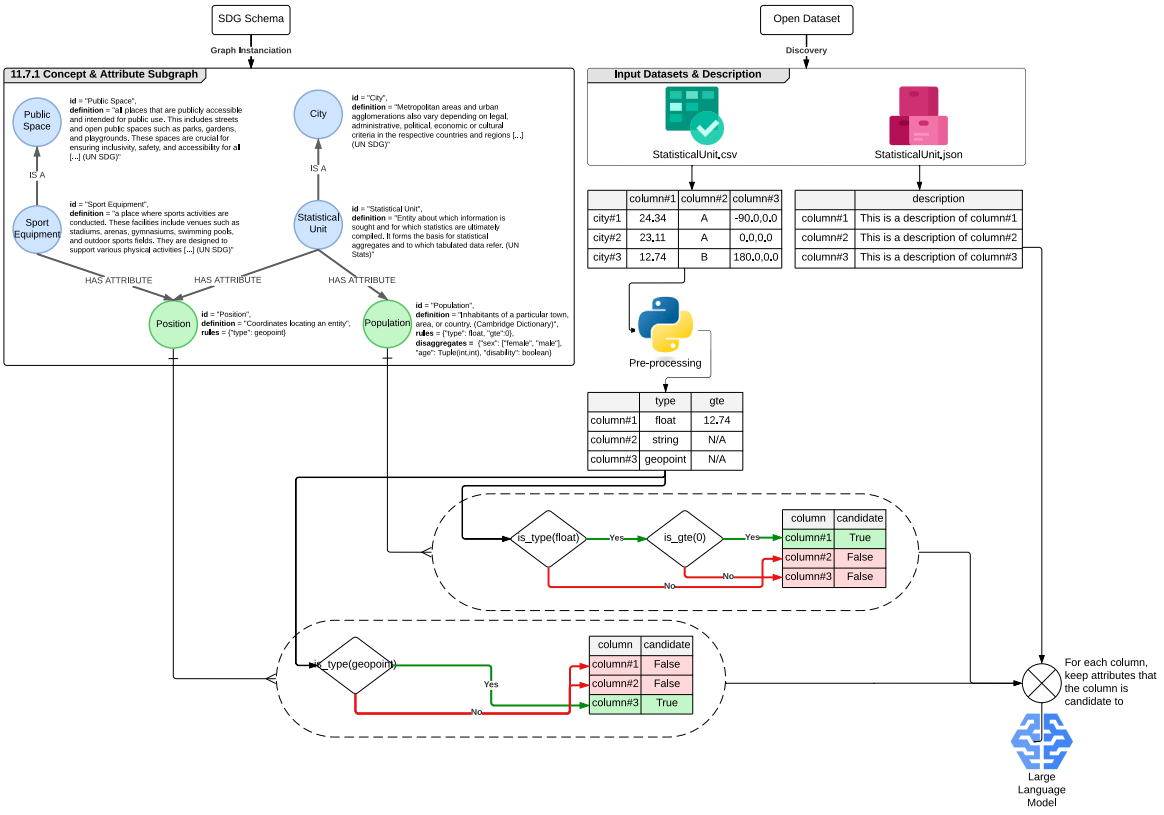

Fig. 9. Driving example process.

are eligible, their descriptions are added to the mapping prompt. The LLM is then asked to choose the corresponding attribute for each attribute's description.

The process in Fig. 9 starts with input datasets, such as *StatisticalUnit.csv* and *StatisticalUnit.json*, containing various columns with data types (e.g., float, string, geopoint). For each column, specific rules are applied to determine its compatibility. For example, column#1 is a candidate because it meets the type requirement (float) and the value condition (greater than or equal to 0). The process involves checking each column against predefined rules like *is_type* and *is_gte*. Next, the LLM is employed to map the columns to the corresponding attributes in the target schema.

To illustrate the approach an example is presented in Fig. 10 where different mapping possibilities are listed and evaluated based on accuracy. The categories for each attribute are as follows:

- **True Negative:** An attribute that does not need to be mapped and is correctly excluded from the mapping process. For example, *FNSAL15P* is a dataset-specific attribute with no corresponding SDG attribute, and it was accurately left unmapped.
- **True Positive:** An attribute that needs to be mapped and is correctly matched to an SDG attribute. For instance, *H0014* is related to population data and was correctly mapped to the Population SDG attribute.
- **False Negative:** An attribute that should have been mapped but was mistakenly left out. For example, the *Coordonnées* attribute represents a position and should have been classified as Position, but it was not mapped.
- **False Positive:** An attribute that was incorrectly mapped to an SDG attribute, even though it is irrelevant for mapping. For instance, *ACTOCC1524* was mistakenly mapped to Population, though it does not correspond to this SDG attribute.

The disaggregation table in Fig. 10 illustrates how attributes are categorized based on their alignment with SDG disaggregation requirements.

- **True:** Successfully mapped to SDG attributes and disaggregated according to pre-defined dimensions. For instance, an attribute such as *H0014* may be categorized as "True" if it aligns with the Population SDG indicator and includes disaggregation dimensions such as sex ("male"), age ([0, 14]), and disability status.
- **False:** Either incorrectly mapped or lack the necessary disaggregation dimensions, leading to misalignment with SDG requirements. For example, an attribute like *POP* may be categorized as "False" if it should have been disaggregated as a Population attribute but was not correctly classified or mapped. This indicates a missed or inaccurate mapping in relation to the intended SDG alignment.

Table 2
Overview of functions in the schema mapping algorithm.

Function	Parameters	Description
`InitializeGraph`	S_1, S_2	Initializes an empty Knowledge Graph $\mathcal{K}_G$ to store the mapped entities and relationships from the source schema S_1 and target schema S_2.
`ExtractDescriptions`	Schema (S)	Extracts descriptions of elements (e.g., entities or attributes) from a given schema. Used to obtain descriptive information that assists in schema mapping.
`ApplyFilteringRules`	Attribute ($\mathcal{A}_S$)	Applies predefined filtering rules (e.g., regular expressions, type consistency) to filter potential matches for the attribute from the source schema. Returns a list of candidate attributes for mapping.
`Compare`	LLM ($\mathcal{F}$), Attribute from source ($\mathcal{A}_S$), Attribute from target ($\mathcal{A}_T$)	Uses the LLM $\mathcal{F}$ to compare an attribute from the source schema with an attribute from the target schema and calculates a similarity score.
`TopKMatches`	Matches ($\mathcal{M}_i$), Attribute ($\mathcal{A}_T$), Score ($\mathcal{M}_{score}$)	Selects the top-k matches for a given attribute, based on the similarity score, ensuring that only the most relevant matches are retained.
`EnrichGraph`	Knowledge Graph ($\mathcal{K}_G$), Match set $\{(\mathcal{A}_S, \mathcal{M}_i, \mathcal{M}_{score})\}$	Adds the identified matches (attributes and their relationships) along with their similarity scores to the Knowledge Graph $\mathcal{K}_G$.
`IsDisaggregable`	Attribute ($\mathcal{A}_S$)	Checks whether a given attribute can be disaggregated into finer sub-categories (e.g., age, location). Returns a boolean value indicating disaggregation potential.
`CheckDisaggregation`	LLM ($\mathcal{F}$), Source attribute ($\mathcal{A}_S$), Matched attribute ($\mathcal{M}_i$)	Evaluates the alignment of disaggregated data between the source attribute and matched target attribute. Returns a disaggregation score indicating the relevance of the disaggregated match.
`EnrichGraphWithDisaggregation`	Knowledge Graph ($\mathcal{K}_G$), Disaggregated match set $\{(\mathcal{A}_S, \mathcal{M}_i, \mathcal{M}_{score})\}$	Incorporates disaggregated matches into the Knowledge Graph $\mathcal{K}_G$, linking finer-grained data relationships.

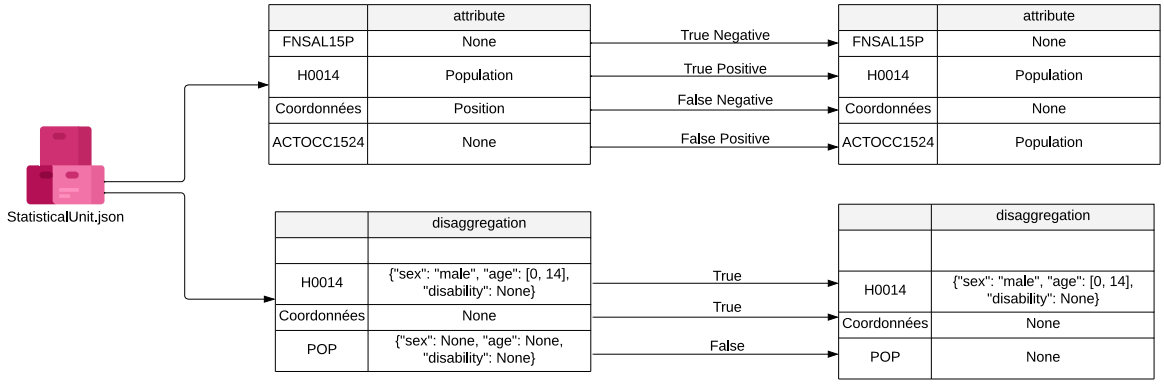

Fig. 10. Disaggregation example.

6.2. Evaluation method

We used two methods for evaluating our approach: direct comparison with existing schema matching methods and an ablation study. By using these evaluation methods, we aim to demonstrate the effectiveness of our rule-based filtering and LLM-powered schema mapping approach.

6.2.1. Comparison with existing mapping methods

Our method's performance was evaluated against Cupid [33], COMA Schema-based [42], COMA Instance-based [43], Similarity Flooding [44], Distribution-based Matching [45] and Jaccard–Levenshtein Matcher [42], which are previously state-of-the-art (SOTA) non-LLM models. Each method has its unique approach and modifications were made where necessary to ensure fair comparisons. Additionally, hyperparameters for each method were tuned to achieve optimal performance, ensuring a robust evaluation of our approach.

We also incorporated few-shot prompting. We manually selected a subset of data instances from the datasets and labeled them to condition the LLM for schema mapping tasks. For our approach, we used two LLM models: GPT-3.5-turbo and GPT-4-turbo. We

Table 3
Grid search results of schema matching methods with their parameters and evaluation metrics.

Entity	Method	Parameters	Accuracy	Precision	Recall	F1-score
SportFacility	Cupid [33]	w_struct: 0.9, leaf_w_struct: 0.01, th_accept: 0.1, few-shot: true	6.19	0.93	100.00	1.85
	DistributionBased [45]	threshold1: 0.01, threshold2: 0.01, few-shot: true	99.12	NaN	0.00	0.00
	JaccardDistanceMatcher [42]	threshold_dist: 0.5, distance_fun: StringDistanceFunction.Levenshtein, few-shot: true	100.00	100.00	100.00	100.00
	SimilarityFlooding [44]	coeff_policy: inverse_product, formula: basic, few-shot: true	99.12	NaN	0.00	0.00
	Our approach	**model: gpt-4-turbo, temperature: 0.0, pre-filtering: True, few-shot: True**	**100.00**	**100.00**	**100.00**	**100.00**
StatisticalUnit	Cupid [33]	w_struct: 0.9, leaf_w_struct: 0.01, th_accept: 0.1, few-shot: true	23.04	17.22	44.83	24.88
	DistributionBased [45]	threshold1: 0.01, threshold2: 0.01, few-shot: true	71.57	NaN	0.00	0.00
	JaccardDistanceMatcher [42]	threshold_dist: 0.5, distance_fun: StringDistanceFunction.DamerauLevenshtein, few-shot: true	82.84	76.74	56.90	65.35
	SimilarityFlooding [44]	coeff_policy: inverse_product, formula: basic, few-shot: true	71.57	NaN	0.00	0.00
	Our approach	**model: gpt-4-turbo, temperature: 0.1, pre-filtering: True, few-shot: True**	**99.51**	**98.31**	**100.00**	**99.15**

perform a grid search for these methods and datasets at hand as shown in Table 3 in order to discover the parameter values resulting in the best performance.

6.2.2. Ablation study

We conducted an ablation study to evaluate the impact of various components of our approach on the performance of the schema mapping. We tested our two models, GPT-3.5-turbo and GPT-4-turbo, with and without pre-filtering, and with and without few-shot prompting. This allowed us to isolate the effects of these components and better understand their contributions to the overall performance. For each evaluation method, we constructed dataframes for the ground truth and predictions.

6.2.3. Evaluation metrics

Accuracy, Precision, Recall, and F1 Score are essential metrics for evaluating a model's performance. They consider both false positives and false negatives, offering a nuanced understanding of a model's predictive capabilities.

Accuracy measures the proportion of correct predictions made by the model across the entire dataset. Precision measures how many of the predicted positive instances are actually positive, while recall measures how many of the actual positive instances are correctly predicted by the model. F1 Score balances precision and recall. It is calculated as the harmonic mean of precision and recall. F1 Score is useful when seeking a balance between high precision and high recall, as it penalizes extreme negative values of either component.

Once we have the manual mappings, we compute the values of true positives (TP), which are the number of correct mappings returned by our method, false positives (FP), which are the number of values that were returned as a true match but should not have been and false negatives (FN), which are the number of values that were not returned as a match but should have been. Additionally, we compute true negatives (TN), which are the number of values correctly identified as non-matches.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.3. Results and discussions

This section presents the results and discussions of our evaluation, including comparisons with existing schema matching methods, an ablation study to assess the impact of some parameters, and a detailed analysis of the performance metrics.

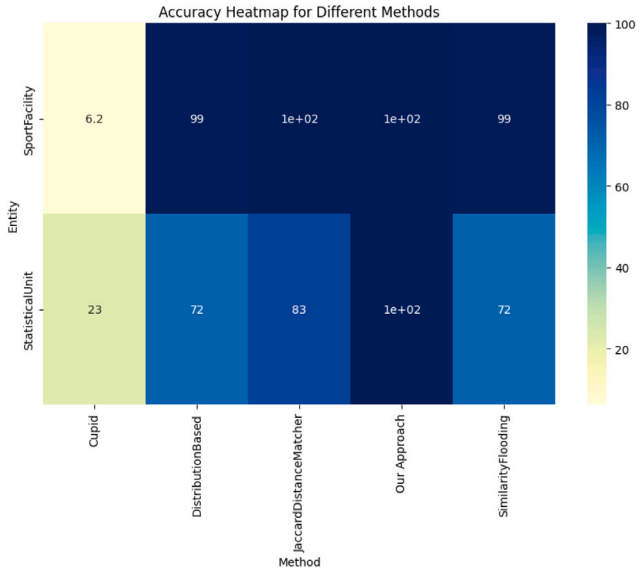

Fig. 11. Accuracy heatmap for different methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.3.1. Comparison with existing mapping methods

Performance comparisons are presented in Table 3. For both studied entities, our approach generally surpasses those of traditional schema matching methods in terms of accuracy, precision, recall, and F1-score.

For the SportFacility entity, the goal is to map well only one attribute which is the position. Our approach using GPT-4-turbo with pre-filtering and few-shot learning achieved perfect scores (100%) in all metrics. This indicates a great identification and mapping of relevant schema elements. In contrast, Cupid achieved only 6.19% accuracy, and Similarity Flooding showed zero precision and recall despite high accuracy. These results suggest that while some methods may return many positive matches, they lack precision and recall. The JaccardDistanceMatcher also performed well, achieving perfect scores, indicating its effectiveness in certain contexts. However, it may be limited by reliance on specific distance functions like Levenshtein.

For the StatisticalUnit entity, the goal is to map a larger number of attributes. Our approach showed high performance with 99.51% accuracy, 98.31% precision, 100% recall, and a 99.15% F1-score. This confirms the robustness of our method. Cupid's performance improved compared to the SportFacility entity but still lagged behind with 23.04% accuracy and a 24.88% F1-score. DistributionBased and Similarity Flooding methods showed inadequate performance with several metrics being zero or NaN. A recall of zero means that none of the relevant matches were identified by the method. In other words, the method failed to retrieve any of the true positive matches present in the dataset. This indicates that the method was unable to correctly identify any of the actual positive cases, resulting in poor performance for recall.

The heatmap shown in Fig. 11 visualizes the accuracy of different schema matching methods across the SportFacility and StatisticalUnit entities. The color intensity represents the accuracy, with darker shades indicating higher accuracy. Our approach and JaccardDistanceMatcher achieved the highest accuracy (100%) for both entities, indicated by the darkest blue cells. Cupid had the lowest accuracy for both entities, with 6.2% for SportFacility and 23% for StatisticalUnit, shown by the lightest cells. DistributionBased and Similarity Flooding had moderate accuracies around 72%. This highlights the superior performance and consistency of our approach compared to traditional methods.

6.3.2. Ablation study

The ablation study results, shown in Table 4, highlight the impact of pre-filtering and few-shot learning on the performance of the GPT-3.5-turbo and GPT-4-turbo models.

For the GPT-3.5-turbo model, the inclusion of few-shot learning significantly improves performance. Without pre-filtering and few-shot learning, the model achieves an accuracy of 21.08%. Adding few-shot learning increases the accuracy to 47.55%. When both pre-filtering and few-shot learning are applied, the accuracy further improves to 69.61%, indicating the combined benefits of these techniques.

The GPT-4-turbo model demonstrates even more substantial improvements. Without pre-filtering and few-shot learning, the model starts with a high accuracy of 67.16%. Adding few-shot learning alone boosts the accuracy to 91.18%. The combination of pre-filtering and few-shot learning leads to the highest accuracy of 99.02%. This shows a marked increase from the baseline, with a significant boost in precision (96.67%) and F1-score (98.31%).

Overall, these results indicate that both pre-filtering and few-shot learning significantly enhance the performance of GPT-based models in schema matching tasks. GPT-4-turbo is 1000 times larger in size reaching 170 Trillion parameters compared to 175 Billion

Table 4
Ablation study metric results.

Model	Pre-filtering	Few-shot	Accuracy	Precision	Recall	F1-score
gpt-3.5-turbo	False	False	21.08	21.86	68.97	33.20
		True	47.55	35.15	100.00	52.02
	True	False	50.49	36.48	100.00	53.46
		True	69.61	48.33	100.00	65.17
gpt-4-turbo	False	False	67.16	46.34	98.28	62.98
		True	91.18	76.32	100.00	86.57
	True	False	87.25	69.51	98.28	81.43
		True	**99.02**	**96.67**	**100.00**	**98.31**

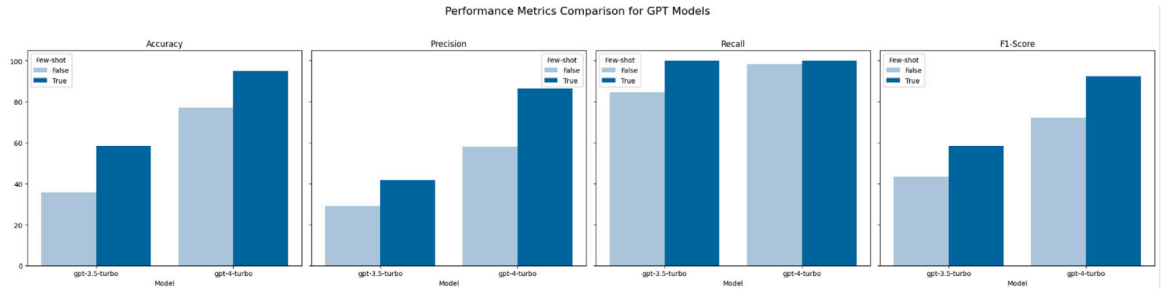

Fig. 12. Performance metrics comparison for GPT models.

parameters of GPT-3.5-turbo. The context length in GPT-4-turbo is 4 to 16 times greater than GPT-3.5-turbo. The context length in GPT3.5 is 2048 and has increased to 8192 and 32 768 (depending on the version) in GPT-4-turbo [46]. Therefore the improvements are more pronounced in the GPT-4-turbo model, demonstrating its superior capability to leverage these techniques for achieving higher accuracy and reliability (see Fig. 12).

Table 4 shows the impact of temperature, pre-filtering, and few-shot learning on the F1-scores of GPT-3.5-turbo and GPT-4-turbo models.

In Table 2, the temperature parameter is adjusted to control the LLM's response style. Typically ranging from 0.0 to 1.0 (with some models supporting values above 1.0 for greater randomness), lower temperatures produce more predictable outputs. Setting the temperature to 0.0 in the first case ensures precise, consistent responses, ideal for factual tasks where accuracy is essential. In the second case, the temperature is set to 0.1, adding a touch of flexibility. This slightly higher setting allows for subtle variations while still maintaining accuracy, making it suitable for tasks that benefit from minor interpretative nuances without sacrificing precision.

GPT-4-turbo with pre-filtering and few-shot learning achieves near-perfect F1-scores across all temperatures. This highlights the effectiveness of these techniques together. GPT-3.5-turbo with both enhancements also performs well, though slightly lower, staying above 80%.

Without pre-filtering and few-shot learning, both models perform worse. GPT-4-turbo drops significantly, and GPT-3.5-turbo falls to around 33%. This shows the importance of these techniques.

In summary, combining pre-filtering and few-shot learning is crucial for high F1-scores, with GPT-4-turbo.

6.4. Indicator calculation

Using the results mapped with LLMs and integrated into our Knowledge Graph, we can calculate specific indicators. An extract f the 2 graphs after the mapping is given in Fig. 13.

The figure shows an extract of the KG linking SDG indicator to relevant data sources, organized into an SDG Graph and a Dataset Graph. The SDG Graph represents entities like *Goal 11* and*Target 11.2*, linked through hierarchical relationships (e.g *HAS TARGET, HAS INDICATOR*), while the Dataset Graph displays data sources, such as *INSEE* and *IGN*, along with attributes like *Population* and *Position*. Thanks to the LLM response, each attribute is mapped to SDG indicators via *IS MAPPED TO* relationships, with mapping scores that reflect the accuracy of each alignment.

Using this KG setup (Fig. 13), we can calculate specific SDG indicators by leveraging the structured mappings between SDG attributes and data sources. For instance, in the case of the french city of `Villejuif`, the proportion of the population with convenient access to sport facilities can be determined by combining population data from INSEE with geographic data from IGN, based on the mapped relationships in the Dataset Graph.

This calculation uses the previously predefined mapping and business rules, such as the accessibility distance threshold of 0.3 km, to filter relevant facilities and population groups. By querying the KG, we can retrieve population counts by age group and check their proximity to mapped sport facilities. Currently, the calculation of this indicator is partially automated through the structured mappings in the KG, but we plan to fully automate this process with a program as part of future work.

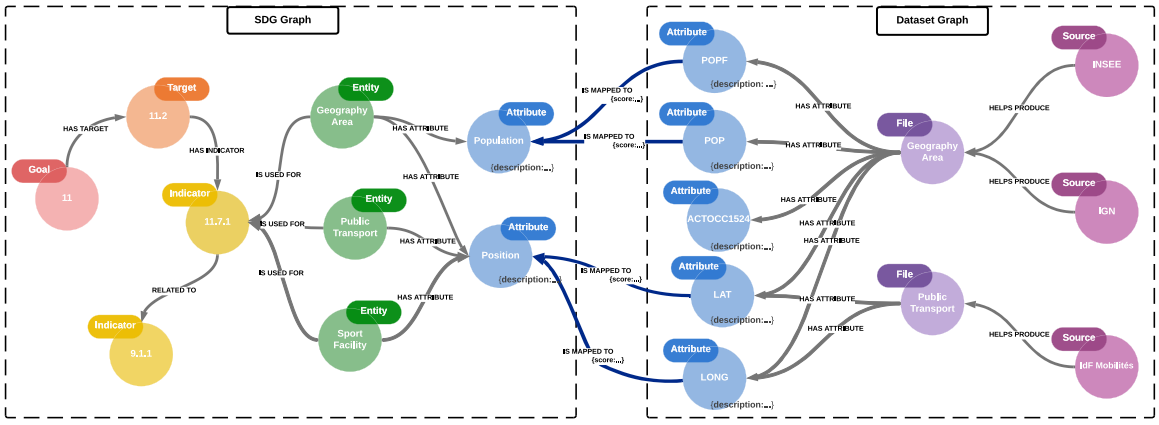

Fig. 13. KG extract illustrating the mapping of SDG indicators to data sources.

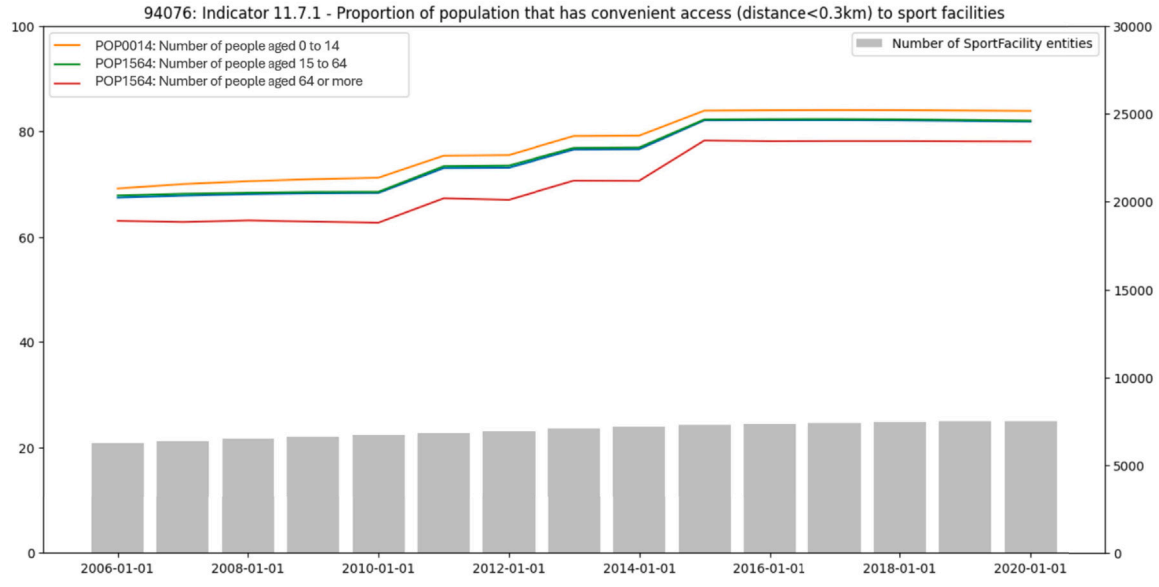

Fig. 14. Proportion of population with convenient access to sport facilities (Distance <0.3 km) by age group and number of sport facilities from 2006 to 2020 in Villejuif (France). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The calculation of this indicator is made possible by our approach, which leverages mapped attributes, a defined structure, and established rules in the KG. This approach supports the calculation process by organizing and aligning the necessary data attributes

Fig. 14 illustrates the proportion of the population with convenient access (distance <0.3 km) to sport facilities for different age groups over time. The x-axis represents the years from 2006 to 2020, while the y-axis on the left shows the percentage of the population, and the y-axis on the right shows the number of sport facilities. The orange line indicates the number of people aged 0 to 14 (POP0014), the green line shows the number of people aged 15 to 64 (POP1564), and the red line represents the number of people aged 65 and older (POP65P). The gray bars illustrate the number of sport facilities.

From 2006 to 2010, there is a slight decrease in accessibility for seniors (POP65P), while adolescents (POP0014) experience a slight increase. Overall, accessibility remains relatively stagnant during this period. From 2010 to 2015, accessibility increases for all population groups, with a noticeable improvement. The total accessibility rises from 68% to 82%, adolescents from 71% to 83%, and seniors from 62% to 78%. Despite these improvements, seniors consistently have the lowest accessibility. From 2015 to 2020, accessibility levels off again, showing little change.

The increase in accessibility between 2010 and 2015 could be linked to the development of transport infrastructure in `Villejuif`, such as the introduction of bus 293 in 2011 and the tramway 7 in 2013. Additionally, the number of sport facilities shows an increasing trend over time, which likely contributes to the improved accessibility.

7. Conclusion

Accurately measuring Sustainable Development Goal (SDG) through their indicators is crucial for guiding global policies toward sustainable development. None of the current methods fully address the generic computation of indicators from open data. Traditional methods often fail to link heterogeneous data sources to specific SDG indicators.

The metadata describing both SDG indicators and open data are available in unstructured textual documents that have to be automatically analyzed to establish correspondences between them. For that, Large Language Models (LLMs) have made significant strides in tackling these tasks, demanding a profound understanding of semantics. However, LLMs present a lack of domain-specific knowledge that can be supplemented using Knowledge Graphs (KGs). The KG's topology allows for effective querying and computation.

By integrating LLMs with KGs, we can enrich the graph with relationships extracted from diverse data sources, aiding in schema mapping and improving the accuracy of indicator computation. This combined approach leverages the semantic understanding of LLMs and the structured data of the KG. All these recent technique have not yet been applied to the studied domain because of the lack of structure in both open data and indicator documentation.

Moreover, to improve LLM performance, we suggest using filtering rules during pre-processing. These rules include regular expressions for pattern matching, type checks to ensure data consistency, and value range validation to filter out anomalies. By incorporating disaggregation rules, we ensure that the LLM can identify and align not only the primary attributes but also their constituent parts

We have tested our approach on SDG Indicator 11.7.1, which measures public space accessibility. Our method showed significant improvements in accuracy, precision, recall, and F1-score compared to traditional methods.

Future work will improve indicator coverage and attribute analysis. We refine filtering rules for better precision. We also aim to establish inference rules on the graph to enable advanced analyses and decision-making. These efforts will support global SDG achievement.

The calculation of this indicator is made possible by our approach, which leverages mapped attributes, a defined structure, and established rules within the Knowledge Graph. This approach supports the calculation process by organizing and aligning the necessary data attributes, but it cannot directly handle raw input data on its own.

CRediT authorship contribution statement

Wissal Benjira: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Faten Atigui:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Bénédicte Bucher:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Malika Grim-Yefsah:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Nicolas Travers:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] U. Nations, Transforming our world: The 2030 agenda for sustainable development, 2015.
[2] H. Guo, D. Liang, Z. Sun, F. Chen, X. Wang, J. Li, L. Zhu, J. Bian, Y. Wei, L. Huang, Y. Chen, D. Peng, X. Li, S. Lu, J. Liu, Z. Shirazi, Measuring and evaluating SDG indicators with big earth data, Sci. Bull. 67 (17) (2022) 1792–1801, http://dx.doi.org/10.1016/j.scib.2022.07.015.
[3] J.-P. Cling, C. Delecourt, Interlinkages between the sustainable development goals, World Dev. Perspect. 25 (2022) 100398, http://dx.doi.org/10.1016/j.wdp.2022.100398.
[4] E. Fotopoulou, I. Mandilara, A. Zafeiropoulos, C. Laspidou, G. Adamos, P. Koundouri, S. Papavassiliou, SustainGraph: A knowledge graph for tracking the progress and the interlinking among the sustainable development goals' targets, Front. Environ. Sci. 10 (2022) http://dx.doi.org/10.3389/fenvs.2022.1003599.
[5] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.-C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (4) (2021) http://dx.doi.org/10.1145/3447772.

[6] A. Joshi, L.G. Morales, S. Klarman, A. Stellato, A. Helton, S. Lovell, A. Haczek, A knowledge organization system for the united nations sustainable development goals, in: R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2021, pp. 548–564.

[7] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Trans. Knowl. Data Eng. (2024) 1–20, http://dx.doi.org/10.1109/tkde.2024.3352100.

[8] M. Howells, S. Hermann, M. Welsch, M. Bazilian, R. Segerström, T. Alfstad, D. Gielen, H. Rogner, G. Fischer, H. Van Velthuizen, et al., Integrated analysis of climate change, land-use, energy and water strategies, Nature Clim. Change 3 (7) (2013) 621–626.

[9] D.K. Joshi, B.B. Hughes, T.D. Sisk, Improving governance for the post-2015 sustainable development goals: scenario forecasting the next 50 years, World Dev. 70 (2015) 286–302.

[10] P. Kumar, F. Ahmed, R.K. Singh, P. Sinha, Determination of hierarchical relationships among sustainable development goals using interpretive structural modeling, Environ. Dev. Sustain. 20 (2018) 2119–2137.

[11] N.A. Almannaei, M.S. Akhter, A. Shah, Improving environmental policy-making process to enable achievement of sustainable development goals, Environ. Policy Law 50 (1–2) (2020) 47–54.

[12] C. Allen, G. Metternicht, T. Wiedmann, National pathways to the sustainable development goals (SDGs): A comparative review of scenario modelling tools, Environ. Sci. Policy 66 (2016) 199–207.

[13] D.J. Abson, J. Fischer, J. Leventon, J. Newig, T. Schomerus, U. Vilsmaier, H. Von Wehrden, P. Abernethy, C.D. Ives, N.W. Jager, et al., Leverage points for sustainability transformation, Ambio 46 (2017) 30–39.

[14] T. Arnold, J.H. Guillaume, T.J. Lahtinen, R.W. Vervoort, From ad-hoc modelling to strategic infrastructure: A manifesto for model management, Environ. Model. Softw. 123 (2020) 104563.

[15] G. Pereira., A. González., G. Blanco., Complexity measures for the analysis of SDG interlinkages: A methodological approach, in: Proceedings of the 6th International Conference on Complexity, Future Information Systems and Risk - COMPLEXIS, SciTePress, INSTICC, 2021, pp. 13–24, http://dx.doi.org/10.5220/0010374600130024.

[16] F. Hanani, S. Aziz, Improving traffic congestion assessment by using fuzzy logic approach, J. Theor. Appl. Inf. Technol. 99 (3) (2021) 625–638.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT-related research and perspective towards the future of large language models, Meta-Radiol. 1 (2) (2023) 100017, http://dx.doi.org/10.1016/j.metrad.2023.100017.

[19] S. Mirchandani, F. Xia, P.R. Florence, B. Ichter, D. Driess, M.G. Arenas, K. Rao, D. Sadigh, A. Zeng, Large language models as general pattern machines, 2023, arXiv arXiv:2307.04721. URL https://api.semanticscholar.org/CorpusID:259501163.

[20] A. Radford, K. Narasimhan, Improving language understanding by generative pre-training, 2018, URL https://api.semanticscholar.org/CorpusID:49313245.

[21] A. Neelakantan, T. Xu, R. Puri, A. Radford, J.M. Han, J. Tworek, Q. Yuan, N.A. Tezak, J.W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T.E. Nekoul, G. Sastry, G. Krueger, D.P. Schnurr, F.P. Such, K.S.-K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, L. Weng, Text and code embeddings by contrastive pre-training, 2022, arXiv arXiv:2201.10005. URL https://api.semanticscholar.org/CorpusID:246275593.

[22] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning–based text classification: A comprehensive review, ACM Comput. Surv. 54 (3) (2021) http://dx.doi.org/10.1145/3439726.

[23] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune BERT for text classification? in: M. Sun, X. Huang, H. Ji, Z. Liu, Y. Liu (Eds.), Chinese Computational Linguistics, Springer International Publishing, Cham, 2019, pp. 194–206.

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

[25] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, B. Jiao, Y. Zhang, X. Xie, On the robustness of ChatGPT: An adversarial and out-of-distribution perspective, 2023, http://dx.doi.org/10.48550/arXiv.2302.12095.

[26] B. Hättasch, M. Truong-Ngoc, A. Schmidt, C. Binnig, It's AI match: A two-step approach for schema matching using embeddings, in: AIDB@VLDB, 2020.

[27] Z. Jiang, F.F. Xu, J. Araki, G. Neubig, How can we know what language models know? Trans. Assoc. Comput. Linguist. 8 (2020) 423–438, http://dx.doi.org/10.1162/tacl_a_00324, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00324/1923867/tacl_a_00324.pdf.

[28] Z. Zhang, Pretrain-KGEs: Learning knowledge representation from pretrained models for knowledge graph embeddings, 2019.

[29] A. Kumar, A. Pandey, R. Gadia, M. Mishra, Building knowledge graph using pre-trained language model for learning entity-aware relationships, in: 2020 IEEE International Conference on Computing, Power and Communication Technologies, GUCON, 2020, pp. 310–315, http://dx.doi.org/10.1109/GUCON48875.2020.9231227.

[30] R. Han, T. Peng, B. Wang, L. Liu, P. Tiwari, X. Wan, Document-level relation extraction with relation correlations, Neural Netw. 171 (C) (2024) 14–24, http://dx.doi.org/10.1016/j.neunet.2023.11.062.

[31] X. Xie, N. Zhang, Z. Li, S. Deng, H. Chen, F. Xiong, M. Chen, H. Chen, From discrimination to generation: Knowledge graph completion with generative transformer, in: Companion Proceedings of the Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 162–165, http://dx.doi.org/10.1145/3487553.3524238.

[32] Z. Chen, C. Xu, F. Su, Z. Huang, Y. Dou, Incorporating structured sentences with time-enhanced BERT for fully-inductive temporal relation prediction, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 889–899, http://dx.doi.org/10.1145/3539618.3591700.

[33] J. Madhavan, P.A. Bernstein, E. Rahm, Generic schema matching with cupid, in: Proceedings of the 27th International Conference on Very Large Data Bases, VLDB '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 49–58.

[34] Y. Li, J. Li, Y. Suhara, A. Doan, W.-C. Tan, Deep entity matching with pre-trained language models, Proc. VLDB Endow. 14 (1) (2020) 50–60, http://dx.doi.org/10.14778/3421424.3421431.

[35] J. Zhang, B. Shin, J.D. Choi, J.C. Ho, SMAT: An attention-based deep learning solution to the automation of schema matching, in: L. Bellatreche, M. Dumas, P. Karras, R. Matulevičius (Eds.), Advances in Databases and Information Systems, Springer International Publishing, Cham, 2021, pp. 260–274.

[36] H. Zhang, Y. Dong, C. Xiao, M. Oyamada, Large language models as data preprocessors, 2023.

[37] M. Souibgui, F. Atigui, S. Ben Yahia, S. Si-Said Cherfi, An embedding driven approach to automatically detect identifiers and references in document stores, Data Knowl. Eng. 139 (2022) 102003, http://dx.doi.org/10.1016/j.datak.2022.102003.

[38] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, Machine Learning Mastery, 2020.

[39] D. Lin, P. Pantel, Concept discovery from text, in: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, Association for Computational Linguistics, USA, 2002, pp. 1–7, http://dx.doi.org/10.3115/1072228.1072372.

[40] Y. Liang, X. Liu, J. Zhang, Y. Song, Relation discovery with out-of-relation knowledge base as supervision, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3280–3290, http://dx.doi.org/10.18653/v1/N19-1332.

[41] L.L. Yan, R.J. Miller, L.M. Haas, R. Fagin, Data-driven understanding and refinement of schema mappings, SIGMOD Rec. 30 (2) (2001) 485–496, http://dx.doi.org/10.1145/376284.375729.

[42] H.-H. Do, E. Rahm, Chapter 53 - COMA — A system for flexible combination of schema matching approaches, in: P.A. Bernstein, Y.E. Ioannidis, R. Ramakrishnan, D. Papadias (Eds.), VLDB '02: Proceedings of the 28th International Conference on Very Large Databases, Morgan Kaufmann, 2002, pp. 610–621, http://dx.doi.org/10.1016/B978-155860869-6/50060-3.

[43] S. Maßmann, S. Raunich, D. Aumüller, P. Arnold, E. Rahm, Evolution of the COMA Match System, vol. 814, 2011.

[44] S. Melnik, H. Garcia-Molina, E. Rahm, Similarity flooding: a versatile graph matching algorithm and its application to schema matching, in: Proceedings 18th International Conference on Data Engineering, 2002, pp. 117–128, http://dx.doi.org/10.1109/ICDE.2002.994702.

[45] M. Zhang, M. Hadjieleftheriou, B.C. Ooi, C.M. Procopiuc, D. Srivastava, Automatic discovery of attributes in relational databases, in: Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, SIGMOD '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 109–120, http://dx.doi.org/10.1145/1989323.1989336.

[46] A. Koubaa, GPT-4 vs. GPT-3.5: A concise showdown, 2023.

Wissal Benjira received her M.S. in Data Science and AI at ESILV Engineering School in Paris. Wissal Benjira is currently standing as a Ph.D. student at DeVinci Research Center (DVRC) and the Laboratory of Geographic Information Sciences and Technologies (LASTIG). Her ongoing research revolves around Information Systems for Decision Support and focuses on the exploration of Open Data to develop City Sustainability Indicators.

Under the co-direction of Nicolas Travers and Bénédicte Bucher, and with the co-supervision of Malika Grim-Yefsah and Faten Atigui, Wissal Benjira is pursuing advanced research in Computer Science.