scientific data

OPEN

DATA DESCRIPTOR

An epidemiological knowledge graph extracted from the World Health Organization's Disease Outbreak News

Sergio Consoli[1]✉, Pietro Coletti[1,2], Peter V. Markov[1,3], Lia Orfei[1], Indaco Biazzo[1], Lea Schuh[1], Nicolas Stefanovitch[1], Lorenzo Bertolini[1], Mario Ceresa[1] & Nikolaos I. Stilianakis[1]

The rapid evolution of artificial intelligence (AI), together with the increased availability of social media and news for epidemiological surveillance, is marking a pivotal moment in epidemiology and public health research. By harnessing the capabilities of generative AI, we use an ensemble approach which incorporates multiple Large Language Models (LLMs) to extract useful epidemiological information for analysis from the World Health Organization (WHO) Disease Outbreak News (DONs). DONs is a collection of regular reports on global outbreaks curated by the WHO with the adopted decision-making processes to respond to them. The extracted information is made available in a knowledge graph, referred to as *eKG*, derived to provide a nuanced representation of the public health domain knowledge. We provide an overview of this new dataset and describe the structure of eKG, along with the services and tools used to access and utilize the data that we are building on top. These innovative data resources open altogether new opportunities for epidemiological research, and the analysis and surveillance of disease outbreaks.

Background & Summary

We are at a pivotal moment in the evolution of epidemiology and public health research[1], significantly influenced by the emergence of artificial intelligence (AI) tools[2]. This advancement is expected to fundamentally reshape the landscape of epidemiological studies, the manner in which infectious disease outbreaks are tracked, and our response to them. Yet, processing the vast quantity of unstructured epidemiological datasets poses significant challenges[3,4]. In this paper we focus on the World Health Organization (WHO) Disease Outbreak News (DONs, https://www.who.int/emergencies/disease-outbreak-news/), an official record of outbreaks maintained by the WHO and operational since 1996. This publicly accessible system disseminates information on events related to health emergencies, reported by national authorities and various collaborators[5,6]. The WHO compiles these data into narrative descriptions that detail the specifics of each outbreak. Typically, these narratives pinpoint the affected geographical area, the disease in question or a syndromic occurrence with an undetermined cause, and may include details of the response efforts, such as the status of laboratory testing for confirmation. Despite their considerable information value, DONs have been underutilized in academic research, primarily due to their unstructured, prose-based format, which poses significant barriers to systematic analysis and integration into automated health informatics systems[7].

In response to these challenges, our study employs an ensemble of advanced generative AI techniques, leveraging the strengths of multiple Large Language Models (LLMs)[8,9], to extract and structure relevant epidemiological information from the WHO DONs[10]. The epidemiological information (i.e., disease and pathogen name, involved country, date of event, cases total, and mortality) extracted via this novel approach is made publicly available adopting *FAIR* (*F*indable, *A*ccessible, *I*nteroperable, and *R*eusable) principles and following the paradigms of Linked Open Data (LOD)[11].

[1]European Commission, Joint Research Centre (JRC), Ispra, Italy. [2]Université catholique de Louvain, Institute of Health and Society (IRSS), Brussels, Belgium. [3]London School of Hygiene and Tropical Medicine (LSHTM), London, United Kingdom. ✉e-mail: sergio.consoli@ec.europa.eu

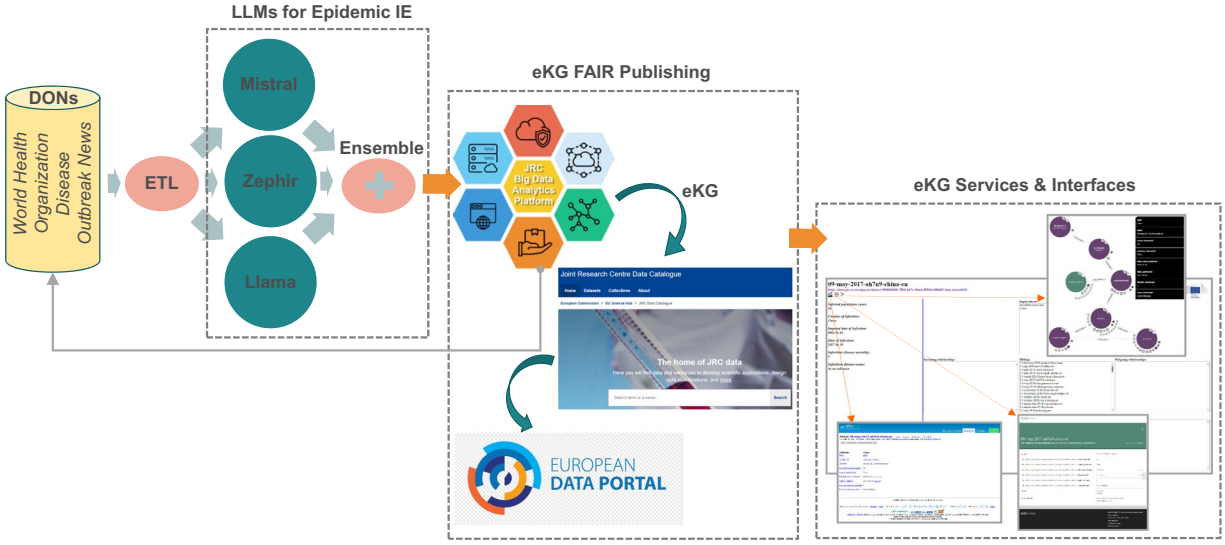

Fig. 1 Schematic overview of the pipeline to extract and publish relevant epidemiological information from the WHO's Disease Outbreak News.

Extracting and integrating heterogenous information for knowledge discovery can be a complex task for researchers and public health experts. To address these issues, we adopt a knowledge-centric paradigm involving the creation of a structured, interconnected, and formal representation of the WHO DONs into a knowledge graph (KG)[12,13], in order to improve the capabilities of conducting complex and extensive analysis of disease outbreaks. In recent years, KGs have become a powerful tool for organizing domain knowledge of various nature and integrating the contained information to reveal hidden associations[14]. They have been successfully used by the research community to create semantically enriched representations of data across various domains[12] as well as to offer significant support to AI systems across multiple domains[15]. KGs are data structures describing the key entities in a domain and their relationships, presenting information in a format accessible and interpretable by both machines and humans[16]. The relationship between two entities is typically formalized as a triple in the format <subject, predicate, object> following the Resource Description Framework (RDF, https://www.w3.org/RDF/) as a standard model[14] (e.g., <*Chikungunya virus, cause, 109 deaths*> or <*Yersinia Pestis, occur, China*>), on top of which different kinds of reasoning activities can be conducted using expressive languages, such as Description Logic (DL)[17], Web Ontology Language (OWL)[18] or SPARQL[19].

While other KG architectures are possible, such as through graph DBMS (such as Neo4j, http://neo4j.org/), each with its own advantages and disadvantages[20], we have chosen to adopt RDF as our data model due to its ability to provide a robust framework for the semantic representation and integration of complex datasets[21]. RDF is ideal for ensuring interoperability and extensibility, as it follows widely accepted standards that enable seamless data integration from diverse sources[22]. This is particularly beneficial in the context of epidemiological data, where the integration of various ontologies and datasets can provide deeper insights and more accurate analysis. Moreover, RDF's compatibility with OWL ontologies allows us to leverage semantic reasoning capabilities, enabling more sophisticated data interpretations and inferencing[18]. These features are essential for enhancing the utility and usability of the produced epidemiological knowledge graph (*eKG*), as they support advanced queries and analysis that go beyond simple data retrieval. By choosing RDF, we ensure that our dataset is not only semantically rich and coherent but also well aligned with the principles of FAIR data, thus promoting broader data sharing and reuse within the scientific and public health communities[11]. By following this approach, we construct the eKG from DONs, offering a nuanced representation of the public health domain of knowledge. This formalized framework effectively integrates extracted epidemiological information from DONs, enhancing the analysis and understanding of public health data.

Figure 1 provides a schematic overview of the pipeline, which will be described in more detail in the "Methods" section. A process is triggered to extract, transform and load (ETL) new reports from WHO DONs, which are then processed by the ensemble of LLMs for the epidemiological information extraction (IE) task. The extracted information is then published using FAIR principles along with the derived eKG. This enables a number of semantic services and interfaces to be deployed above, allowing access and exploration of the derived knowledge graph. These resources have significant potentials to improve the utility and usability of WHO DONs for both researchers and public health professionals, and promote new research opportunities in public health and epidemiology for the analysis of disease outbreaks, surveillance, and response to public health threats.

In summary, this data article makes a contribution in the field of epidemiology by providing a comprehensive and structured dataset derived from the WHO DONs. Our work showcases the value of extracting epidemiological data from narrative reports. The establishment of a knowledge graph of epidemiological information and the application of LLMs in an ensemble fashion[10,23] to process and analyze unstructured data mark a transformative step towards enhancing our global understanding and management of disease outbreaks.

Related Work

For a better understanding of the approach that we have followed in the construction of eKG, we first outline the methods developed for integrating graph-based biomedical heterogeneous data, and then describe the main characteristics of the WHO Disease Outbreak News dataset from which eKG has been extracted.

Methods for developing knowledge graphs in the biomedical domain. The challenge of combining heterogeneous data sources represents a fundamental issue in contemporary data management systems, prompting the development of numerous methodologies for processing relational information[24]. The widespread use of diverse data formats has highlighted the essential role of ontological frameworks[25]. Ontologies can be used to reconciliate heterogeneous information[26] as well as to query the data by overcoming interoperability issues[27]. In this case, the relationships between ontological concepts and data instances are declared via explicit mapping specifications. Afterwards, such declarations are used to build the queries on the data instances respecting the adopted ontological definitions. This can be achieved through virtualization[28], by using dynamic real-time transformations during query execution. Queries translations in this approach are obtained by leveraging mapping specifications and ontological relations. Virtualization involves directly fetching only the data needed for the query from the original repositories. However, it may introduce latency and inconsistencies when underlying source structures undergo modifications[29]. On the other hand, querying ontological data can be also achieved via the materialization method, which synchronizes local data architectures with the semantic models and their associations. Materialization, in contrast to virtualization, offers data retrieval from the ontological system by consolidating the collected information into a unified storage solution, although frequent updates may compromise data quality[29].

Multiple approaches have been developed to address data heterogeneity through mapping rule specification. Prominent examples of such methodologies can be schematized in the following:

- CSVW, a framework for characterizing tabular CSV data through metadata annotations to enable RDF conversion[30];
- SPARQL-Generate, which augments SPARQL functionality to facilitate data transformation from varied formats into RDF by incorporating RDF generation capabilities alongside SPARQL query processing[31];
- R2RML, a W3C recommendation for transforming relational database content to RDF format[32];
- RML, which extends R2RML functionality to accommodate diverse data structures[33];
- ShExML, a comprehensive mapping language that merges RDF Shape Expression (ShEx) characteristics with transformation capabilities to enable conversion of multiple data formats into RDF structures[34];
- YARRRML, a human-friendly, textual representation of RML that streamlines RDF mapping rule creation by supporting YAML-based mapping specifications[35].

Within the biomedical field, considerable research initiatives are focused on constructing knowledge graphs through the consolidation of numerous publicly available datasets, leveraging both materialization and virtualization methodologies. As an illustration, Zhang *et al.*[36] incorporated fusion and annotation processes to establish an integrated KG that encompasses multiple sources while being enhanced with biological ontological structures, including Disease[37], Gene[38], Plant[39], and Trait[40] ontolologies. In addition, ReproTox-KG[41] aims to predict placenta defects that can be derived from compounds. In particular, this knowledge graph extracts information from peer-reviewed publications and integrates ontological frameworks and taxonomies including the CDC birth-defect terms (https://www.cdc.gov/birth-defects/about/types.html), the Human Phenotype Ontology (HPO)[42], SigCom LINCS datasets[43] for drug-gene interaction networks, DrugCentral[44] for mapping pharmaceutical compounds to developmental defect categories, and Geneshot[45] for establishing gene-to-birth-defect associations. Instead, the BioGrakn KG project[46] integrates genomic, proteomic, and clinical data using the capabilities of the Grakn's knowledge graph (https://grakn.ai/) to support complex querying and inference, drawing on sources like UniProt[47] and the Gene Ontology[38]. The MONARCH Initiative[48] unifies cross-species genetic, genomic, and phenotypic data to explore genotype-phenotype relationships, incorporating data from resources such as the Human Phenotype Ontology (HPO)[42] and the Mouse Genome Informatics (MGI) database[49]. Additional initiatives include the Oregano KG, which focuses on drug repositioning[50], PrimeKG[51], that provides detailed disease insights, and a federated framework[52] designed to facilitate semantic queries across various distributed bioinformatics data sources using ontology-based integration.

The aforementioned research collectively demonstrates the difficulties arising from attempts to merge disparate data repositories characterized by distinct modeling approaches and semantic frameworks. Particular attention is directed toward successfully resolving challenges such as the unavailability of consistent naming conventions, data overlap, and repeated records. In our case, when integrating WHO DONS, we also have to consider the lack of comprehensive ontologies that cover all possible epidemiological reports. Existing ontologies are often fragmented, addressing only certain aspects of a disease outbreak event. These challenges need to be addressed during the creation of the eKG. In the following sections, we will elaborate on the peculiarities of our method for integrating data from epidemiological reports into a graph-based heterogeneous representation. Our approach uses semantic similarity measures to improve the accuracy of entity integration from diverse data, as seen in GOntoSim[53]. Additionally, unlike graph DBMS like Neo4j, our use of RDF provides a more standardized method for organizing interconnected knowledge, facilitating compatibility with existing ontological frameworks. Our approach also leverages the open-source NCBO BioPortal[54] to integrate biomedical ontologies, ensuring consistent and accurate entity representation across datasets by grounding nodes to ontology terms, thereby preventing or minimizing data duplication.

The WHO Disease Outbreak News (DONs). The WHO DONs platform is a critical source of information on verified public health incidents or potential events of concern. Since its launch in January 1996, the platform provided timely updates on various public health incidents. The core focus of DONs is to share news about confirmed or potential public health events, including unexplained incidents with significant international health implications, diseases with known causes that have the potential to spread globally, and events that could disrupt public health interventions, international travel, or trade. This repository of information helps international actors stay informed about health-related events and emergencies around the world, and is also occasionally used by the research community as a public record of outbreak events[4–7,55]. However, the Disease Outbreak News is not an exhaustive record of all known disease outbreaks globally. The criteria that the WHO employs to determine which epidemiological situation updates to release are not entirely clear[4,55]. Over the nearly 30 years since the DON has been utilized, the reports' content and topics may have subtly evolved, especially as reporting obligations have shifted, such as with the adoption of the revised International Health Regulations (IHR, https://www.who.int/health-topics/international-health-regulations) in 2005. Nevertheless, with over 3,000 verified news articles on epidemics published since its launch, DONs remain an official and authoritative source of information on public health events.

The analytical value of WHO DONs is significantly limited by its unstructured, narrative format. Each report is presented on an individual webpage, and although these pages are indexed by search engines like Google, there is little in the way of scaffolding or standardized metadata to facilitate accessing reports through more structured queries. The current interface makes it difficult to quickly retrieve information or conduct analysis. To facilitate global health research that utilizes and exploits DONs as a comprehensive record of outbreaks and a valuable repository of documents, we developed a standardized framework for sharing epidemiological information, as described in the next section, to allow DONs to be more useful to researchers and public health professionals.

Methods

This section describes how we constructed the knowledge graph of epidemiological information from DONs. The released resource was built using the automatic pipeline depicted in Fig. 1, composed of several modules that can be grouped into three main categories: *i) LLMs for Epidemiological Information Extraction (IE)*, where the important epidemiological information is extracted from DONs using an ensemble of open-source LLMs, *ii) eKG FAIR Publishing*, which converts the extracted information into the eKG knowledge graph following the paradigm of Linked Open Data, adds additional contextual information to the extractions, and makes the resource available to local authorities, research organizations, national and European governmental institutions and international organizations through a public repository, and *iii) eKG Services & Interfaces*, consisting of several additional services and tools that we are building on top to expand the final user experience.

LLMs for Epidemic IE. This study utilizes LLMs, a type of artificial intelligence model that leverages deep learning architectures to analyze complex patterns in language data, to extract useful epidemiological information for analysis. These predictions are informed by the model's learning from extensive text datasets. The development of LLMs is based on the Transformer architecture[8], a type of model design used in machine learning and AI that enables the efficient processing of sequential data by using self-attention mechanisms to capture long-range dependencies in input sequences[56]. In contrast to classical sequential models like Long Short-Term Memory (LSTMs) and Recurrent Neural Networks (RNNs), Transformers can be trained in a highly parallelized and efficient manner, while also being particularly effective at capturing long-range dependencies in sequences, making them a powerful tool for natural language processing tasks[8,56].

Building on the assessment of recent LLMs for epidemiological information extraction[10], our work employs an ensemble approach, which was shown to be comparable or even superior to commercial techniques for extracting structured epidemiological information from text, including outbreak disease names, involved countries, event dates, total cases, and deaths[10]. A solution based on open-source LLMs is ideal in this context because it overcomes specific usage constraints of commercial models, e.g. maximum number of tokens that can be processed per day, maximum number of API requests per units of time etc., allowing for a full deployment on the entire DONs data.

By leveraging the extracted information in the context of infectious disease outbreaks from DONs, and integrating it into a surveillance system, we aim to improve the precision and speed of epidemiological analysis. After extracting, transforming and loading (ETL) the new reports from WHO DONs, cleaned and normalized, they are processed by the ensemble of those LLMs. The adopted LLMs for the ensemble are: *Mistral-7B-OpenOrca*, *Zephyr-7B-Beta*, and *Meta-Llama-3-70B-Instruct*. In the following sections we provide a brief description of these models.

Mistral-7B-OpenOrca. The *Mistral-7B-OpenOrca* model (https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca), an open-source variant fine-tuned by OpenOrca on its proprietary datasets (https://huggingface.co/datasets/Open-Orca/OpenOrca), is built above the decoder-only Mistral LLM. Mistral models make use of a significant number of computational advancements with respect to other state-of-the-art LLMs, e.g. sliding-windows attention, which allows the model to extend the number of tokens it can simultaneously process. Despite having a relatively smaller size of 7 billion parameters, *Mistral-7B-OpenOrca* demonstrates strong performance across a range of tasks, while also providing fast prediction time. Notably, its performance is often comparable to larger models, including the Llama 3 models family. Equipped with a maximum length of the context equal to 4,096 tokens, the *Mistral-7B-OpenOrca* model stands out as one of the top-performing open-source models with similar parameter sizes (around 7 billion parameters). It excels particularly in tasks requiring efficient resource usage and fast inference time, making it an attractive choice for various API applications.

Zephyr-7B-Beta. The *Zephyr-7B-Beta* model (https://huggingface.co/HuggingFaceH4/Zephyr-7B-Beta) is an open-source LLM leveraging the transformer architecture. It is a Mistral variant obtained by fine-tuning the model on a collection of synthetic and public datasets. This model employs a unique training approach, utilizing a combination of masked language modeling and next sentence prediction objectives, allowing it to learn a rich representation of language and generate coherent and context-specific text. Similarly to the *Mistral-7B-OpenOrca* model, this model has also 7 billion parameters, which are distributed across 24 layers of transformer blocks, each consisting of self-attention mechanisms and feed-forward neural networks. Notably, this model utilizes a variant of the attention mechanism known as scaled dot-product attention, which enables it to efficiently process long-range dependencies in input sequences. Although being compact in size, it often outperforms larger LLMs in language understanding and generation tasks, demonstrating performance comparable to several models in the 20-30B range. The model is characterised by a context length equal to 4,096 tokens and is quick at inference, making it an attractive option for deployment in resource-constrained environments. The *Zephyr-7B-Beta* model is considered one of the top-performing open-source models with fewer than 30 billion parameters, making it an effective LLM for various API applications.

Meta-Llama-3-70B-Instruct. The *Meta-Llama-3-70B-Instruct* model is a state-of-the-art open-source LLM developed by Meta (https://ai.meta.com/llama/), leveraging the transformer architecture and building upon the success of its predecessors. It stands as one of the leading open-source LLMs on the market, matching the performance standards of the renowned GPT models. The core Llama model was developed using an extensive and heterogenous selection of public online documents, which enables the model to have an in-depth grasp of language patterns and contexts. The fine-tuned *Meta-Llama-3-70B-Instruct* model advances this grasp by incorporating public instructional data and extensive human annotations. This model is fine-tuned on this massive dataset of instructions, which enables it to excel in tasks that require precise and context-specific responses (https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct). The fine-tuning process involves a combination of masked language modeling and next sentence prediction objectives, allowing the model to learn a rich representation of language and generate coherent and context-specific text, ensuring a high level of precision and adaptability in its responses.

The *Meta-Llama-3-70B-Instruct* model boasts 70 billion parameters, which are distributed across 32 layers of transformer blocks, each consisting of self-attention mechanisms and feed-forward neural networks. With a context length of 4,096 tokens, this model is capable of handling complex language tasks, making it an ideal choice for applications that require advanced language understanding and generation capabilities. Furthermore, the model's instruction-based fine-tuning process ensures that it is highly adaptable and can respond accurately to a diverse range of prompts and queries, solidifying its position as a top-tier LLM in the field.

Ensemble. Ensemble learning[57] is a machine learning approach that merges several models to address a specific problem. Unlike conventional machine learning techniques that strive to derive a single hypothesis from training data, ensemble learning aims to build multiple hypotheses and merge them to achieve a more precise and robust outcome[57]. This approach can be applied to various AI paradigms, including those that utilize multiple techniques to address the same task. The *Ensemble* model is an advanced AI strategy that capitalizes on the capabilities of the different LLMs, i.e. *Mistral-7B-OpenOrca*, *Meta-Llama-3-70B-Instruct*, and *Zephyr-7B-Beta*, to generate the final outputs. For a validation of the produced LLMs outputs see the next "Technical Validation" Section. Practically, the *Ensemble* model employs a majority voting mechanism to choose the most precise result from the contributing models. This approach guarantees that the output with the greatest agreement among the models is selected, thus improving inference reliability and precision. In the event of a tie, where each of the three models provides a different answer, the ensemble resolves the situation by selecting one of the answers at random. By combining the strengths of multiple models, the *Ensemble* can produce more robust and accurate results, increasing its confidence inference. The *Ensemble* benefits from the distinct advantages that each model contributes. In particular, *Mistral-7B-OpenOrca* is known for its strong performance across various tasks, earning popularity for being both adaptable and reliable. *Zephyr-7B-Beta*, on the other hand, is recognized for its sophisticated capabilities and top-tier performance. Lastly, *Meta-Llama-3-70B-Instruct* is well-regarded for its proficiency in chat-based interactions, providing high-quality, contextually accurate responses, further enhancing the ensemble's overall effectiveness.

Each LLM is separately launched upon the new epidemiological DONs reports. All LLMs share the same context length of 8K tokens, meaning they can process at one time a text of maximum length of 8K tokens when generating a response or performing a task. To note, the number of tokens corresponding to a certain number of words can vary greatly depending on the language, the tokenization method used, and the specific text. In English, with common tokenization methods, 100 tokens corresponds to approximately 75 words. Therefore, as a rough estimate, 8K tokens would correspond to around 6K words. In the case the DON report to process exceeds this context length value, it is split in pieces which then undergo a preliminary summarization step, using the following prompt:

> "*Summarize the epidemiological text below, focusing on any aspects that are relevant to the disease outbreak, place and time of the infectious disease outbreak occurred, and the number of cases and deaths derived.*
> *Do not invent. Write no explanations or notes.*
> *Text: ...*"

Afterwards, for the extraction of the key epidemiological entities (name of disease outbreaks, involved country, date of the outbreak, number of confirmed cases and deaths) for each DON report, the LLMs are supplied with the prompt depicted below[10]:

"*From the text below extract the following items:*
1 - The name of the disease that caused the outbreak.
2 - The name of the country where this disease outbreak occurred, if present.
3 - The date when this disease outbreak occurred, if present. Show the date in the format YYYY-mm-dd.
4 - The number of deaths caused exclusively by the disease outbreak mentioned in the text, if present.
Format your response as a JSON object with the following keys: disease name, country, date, cases. If the information is not present, do not invent and use "None" as the value.
Text: ..."

The used prompts were selected based on their ability to effectively extract key epidemiological entities with high accuracy, as demonstrated through iterative testing and validation against a benchmark portion of the data. This ensures that the extracted information is both precise and reliable, aligning with the objectives of our study. As an example, consider the WHO DON "Nipah virus - India" of the 31st May 2018 (https://www.who.int/emergencies/disease-outbreak-news/item/31-may-2018-nipah-virus-india-en), reporting a Nipah virus outbreak in Kerala, India, resulting in 15 laboratory-confirmed cases, 13 deaths and 2 hospitalizations, with bats linked to the initial transmission, and prompting a robust local and WHO-supported public health response, including surveillance and infection control measures. The extraction answer provided in JSON by the *Mistral-7B-OpenOrca* model using the detailed prompt was:

{ *"disease": "Nipah virus", "country": "India", "date": "2018-05-19", "cases": "15", "deaths": "13"* },

which we can note correctly structures the main epidemiological information from this disease outbreak.

In addressing the challenge of ensuring valid JSON outputs from LLMs, we opted for an algorithmic validation approach. This involves programmatically verifying JSON validity and employing synthetic heuristics to attempt reconstruction in case of JSON formatting issues. These heuristics involve checking for common JSON structural issues such as typos, missing or extra characters, and unexpected attributes. The process involves:

1. Checking if the output is a valid JSON using a validation function.
2. If invalid, attempting to clean and correct common errors such as misplaced or absent quotation marks, misplaced brackets, and removing irrelevant text.
3. Ensuring all necessary attributes are present and properly formatted.
4. Removing any unexpected attributes not required in the JSON structure.

For the outputs we were unable to clean, which accounted for less than 1% of the overall data from our calculations, we discarded them. This method allows us to effectively manage and correct JSON outputs, ensuring they meet the necessary structural requirements for further elaboration.

After a report is processed by all the contributed LLMs, majority voting is used among their outputs for each of the extracted epidemiological information, forming in this way the output of the *Ensemble*. This majority voting technique is particularly effective in mitigating the hallucination issue commonly associated with LLMs. Hallucination refers to the generation of information that is not present in the input data[58]. In machine learning, ensemble methods aggregating the predictions from multiple approaches are able to improve the overall performance. This is indeed a well-documented phenomenon in machine learning[23,57]. The underlying principle is that by pooling together the strengths and mitigating the weaknesses of various models, the ensemble can achieve better accuracy and robustness than any single model alone. By employing majority voting on the extracted epidemiological information from the LLMs, the ensemble approach ensures that only the most consistent and agreed-upon information is retained, thereby reducing the potential for hallucinated outputs and improving the reliability of the extracted epidemiological data. For example, if the extracted number of cases from the three LLMs for a report is: 15, 17, and 15, the *Ensemble* would give 15 as the output for the number of cases for that report. Please note that if each model had given a different answer, the ensemble randomly selects one. When textual epidemiological information, that is disease names and countries, need to be evaluated by the majority voting system, the process is slightly more complicated, as two terms representing semantically the same concept should be grouped together. This was achieved by compiling a dictionary of synonyms of disease and pathogen names, and another of country names synonyms. Synonyms are created by (i) first checking for syntactic similarity among the terms to be compared (e.g. "Trinidad and Tobago" and "Trinidad & Tobago", or "Florida, USA" and "USA (Florida)"; (ii) checking for synonymy by looking for overlapping synsets in WordNet[59]; (iii) checking for semantic similarity among the terms to be compared (e.g. "Middle East respiratory syndrome" and "MERS-CoV", or "Dengue" and "Breakbone fever". To calculate the semantic similarity we used Sentence Transformers, also known as SBERT (Sentence-BERT)[60], which is a modification of the pre-trained BERT (Bidirectional Encoder Representations from Transformers)[61] network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings. These embeddings can then be compared using a metric like the cosine similarity[62] to determine the semantic similarity between sentences. SBERT fine-tunes the BERT network on a dataset of sentence pairs, training it to produce embeddings that bring semantically similar sentences closer together in the embedding space. This allows for much faster performance in downstream tasks, as the sentence embeddings can be precomputed and easily compared without the need for further BERT processing. In particular, to evaluate the semantic similarity among country names, we experimentally selected the *all-mpnet-base-v2*

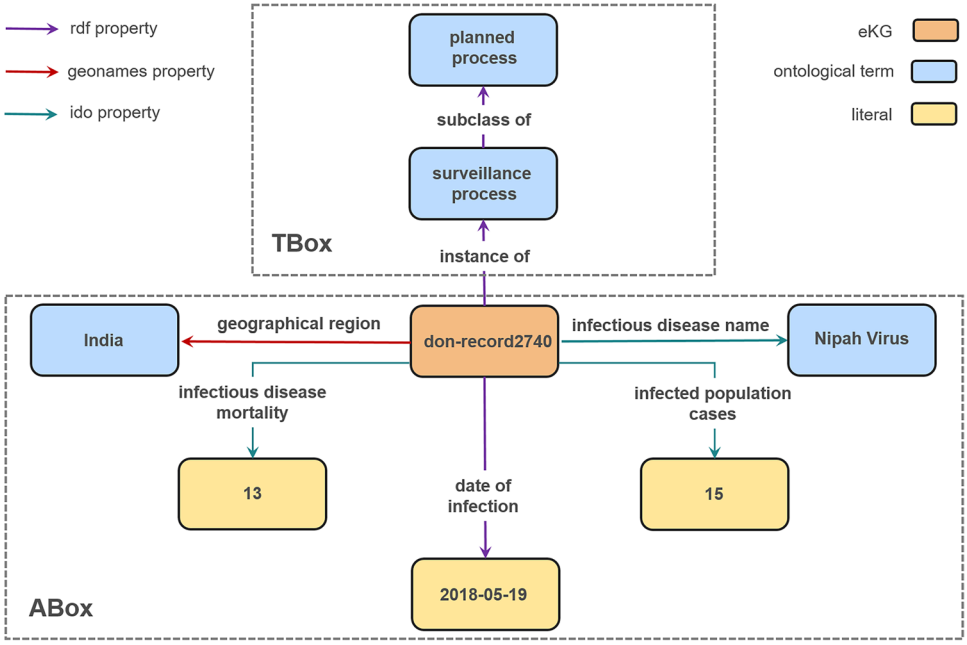

Fig. 2 This picture illustrates an example on the implementation of Description Logic for structured knowledge representation. The TBox comprises concepts such as `surveillance process` and `planned process`, along with assertions defining relationships between these concepts (e.g., `<surveillance process, subclass of, planned process>`). The ABox represents instances of these concepts, like `don-record2740`, and provides assertions relating these instances to other concepts (e.g., `<don-record2740, instance of, surveillance process>` and `<don-record2740, geographical region, India>`) as well as to literal values (e.g., `<don-record2740, date of infection, 2018-05-19>` and `<don-record2740, infected population cases, 15>`).

(https://huggingface.co/sentence-transformers/all-mpnet-base-v2). SBERT model, which is an all-round model trained on a large and diverse dataset of over 1 billion training pairs and tuned for many diverse use-cases. For the disease outbreaks we used instead BioBERT[63], a domain-specific version of the BERT model pre-trained on biomedical literature to enhance performance on biomedical natural language processing tasks[64]. We set an experimental threshold of 0.8 for the semantic similarity $\in [0, 1]$, meaning that two terms having a semantic similarity larger than this threshold were considered to represent the same concept. Choosing the most appropriate semantic similarity score can be challenging and often requires empirical experimentation to determine the best fit for a specific scenario. A typical range for semantic similarity values is often between 0.7 and 0.9, which has been supported by numerous studies (see e.g.[53,65,66]). For our use case, we experimentally selected a threshold of 0.8, as it demonstrated an effective balance in differentiating terms while maintaining homogeneous groups of synonyms for disease and pathogen names, as well as country names. In this way, clusters of disease names and countries synonyms were obtained, forming dictionaries of similar concepts.

Using the majority voting approach, the *Ensemble* model determines the most precise output from the three participating models, thus boosting the chances of accurate and reliable predictions. This approach enables the model to compete with the performance of more advanced commercial language models for epidemiological IE[10]. By combining the advantages of multiple models, it emerges as an effective strategy for the task of epidemiological information extraction.

eKG FAIR Publishing. The epidemiological knowledge graph, eKG, has been produced from the information extracted by the WHO DONs, providing a rich data model for the public health domain. It has been implemented by using Semantic Web technologies, such as RDF (https://www.w3.org/TR/REC-rdf-syntax/) and OWL (https://www.w3.org/OWL/), which allow human experts to verify, curate, and correct both the data and their schema, to produce an ontologically-grounded knowledge graph.

A knowledge graph can be formally defined as a pair `<T, A>`, where `T` is the TBox and `A` the ABox[67]. The TBox serves as the foundational layer containing the formal descriptions of a domain's structure, encompassing specific class hierarchies, relational properties, and basic axioms or declarations[67] (e.g., a `surveillance process` is a subclass of a `planned process`, as illustrated in Fig. 2). In contrast, the ABox details the properties and attributes of instantiated class objects (namely, individuals or data names) and declarations concerning their categorical placement within the TBox structure[67] (e.g., as depicted in Fig. 2, the `don-record2740` is a `surveillance process` associated to the `Nipah Virus` infectious disease occurred in `India` on `2018-05-19`, with confirmed `15` infected population cases and `13` deaths). External

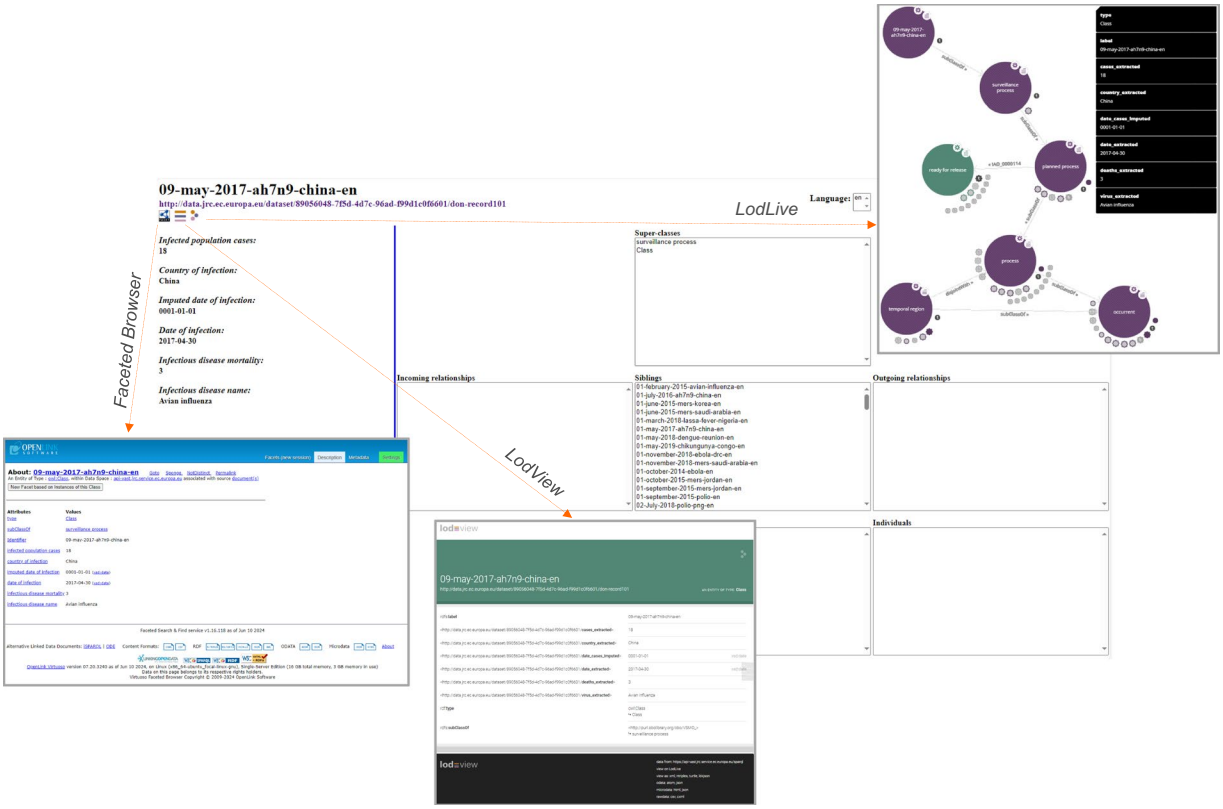

Fig. 3 Snippet of eKG exploration services and interfaces.

data entities that exist beyond ontological specifications can be formulated using existing ontological paradigms through TBox strategies (by leveraging schema-level constructs) or ABox implementations (by applying individual-level representations)[68]. The class-based method characterizes each data element via `subClassOf` relations with existing ontology classes, while the instance-based method characterizes them via `instanceOf` relations with existing ontology classes, as done in eKG. In particular, for the extraction of eKG we have strongly followed principles from ontology design patterns (ODPs)[69] and LOD[11], that is, whenever possible we reuse existing ODPs from online LOD repositories. Reused patterns are annotated with the OPLa framework[70] in order to facilitate class mapping and identification, which helps ground the knowledge graph to the ontology classes, ensuring semantic consistency and enhancing data integration. In particular we reuse classes and relations from IDO (Infectious Disease Ontology)[71] and include explicit alignments to them by `rdfs:subClassOf` properties. IDO is a collection of structured entities generally relevant to both the biomedical and clinical aspects of infectious diseases. The structure of IDO adheres to the Basic Formal Ontology (BFO)[72] and aims to provide a coverage of the infectious disease domain. Terms in IDO that are within the scope of other OBO Foundry ontologies[73] are derived from those respective ontologies. In this way we also map indirectly to the WHO's International Classification of Diseases, Version 10 (ICD10, https://icd.who.int/browse10/2019/en).

When creating a knowledge graph from heterogeneous sources, like in our case with eKG, there is the possibility of introducing the same real-world entity multiple times, as they may be expressed in different forms (e.g. "SARS" vs "severe acute respiratory syndrome"). *Entity linking*[74], i.e. grounding the facts in domain ontologies, is a technique used to reduce this phenomenon and an essential feature of knowledge graphs. In our case, the data resources of eKG have been linked to biomedical ontologies using the NCBO BioPortal (https://bioportal.bioontology.org/), and to geo-locations using GeoNames (https://www.geonames.org/), ensuring consistent and accurate entity representation across datasets.

The NCBO BioPortal[54] is a comprehensive online platform designed to provide access to a vast repository of biomedical ontologies and related resources. Developed by the US National Center for Biomedical Ontology (NCBO), the BioPortal allows researchers, clinicians, and developers to browse, search, and integrate diverse ontological data to enhance biomedical research and application development. It supports a wide array of functions including ontology browsing, download, versioning, and annotation, facilitating better understanding and interoperability of complex biological and medical data. By offering these tools and resources, the BioPortal plays a critical role in advancing semantic technology and data sharing in healthcare and life sciences.

GeoNames, conversely, is an open-source geographical database, regularly updated, that covers all countries worldwide and contains over eleven million place names. We have adopted it to describe geographical entities, such as cities, regions, or country names where a disease outbreak has occurred[75].

Both BioPortal and GeoNames represent invaluable resources for knowledge graph construction in a field such as public health intelligence, where data quality is essential to avoid data duplication and ensure unique

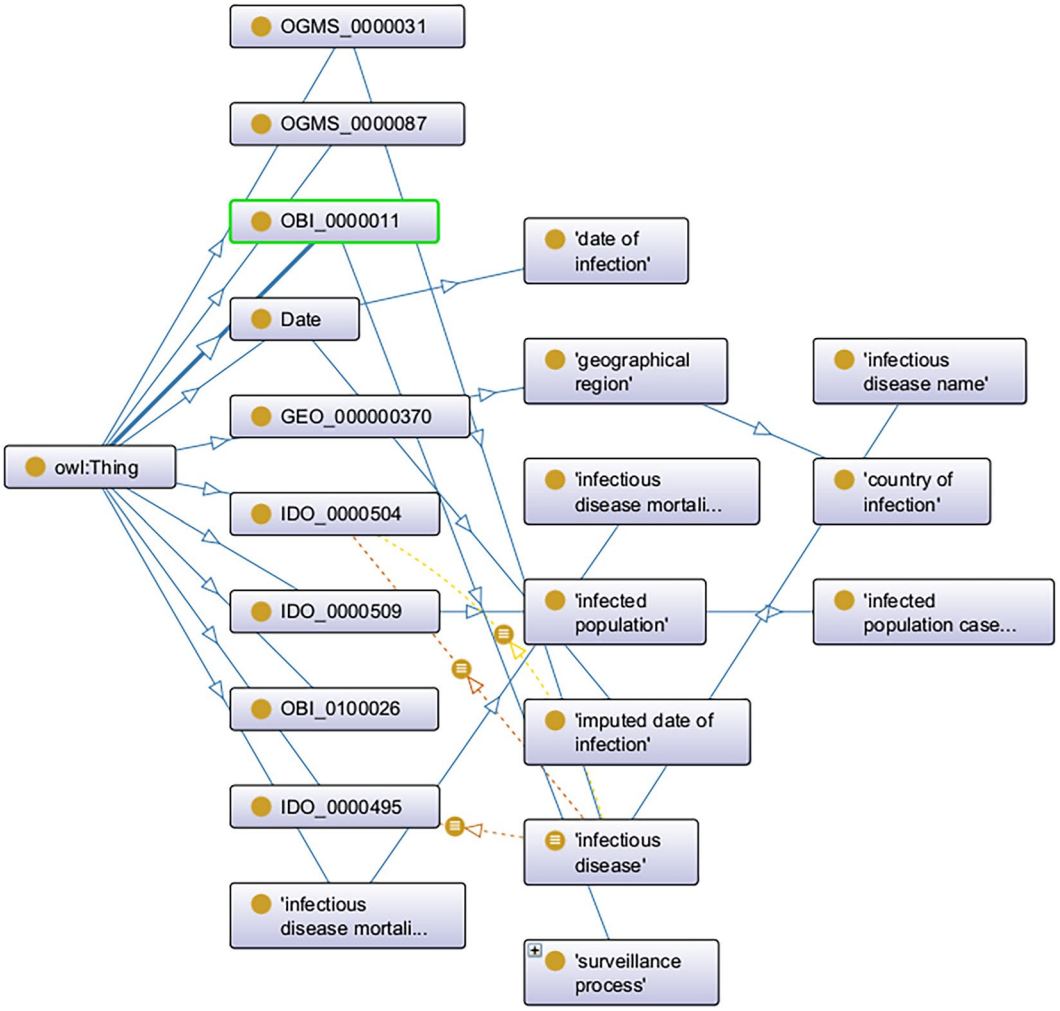

Fig. 4 Main classes within eKG.

identification of outbreak events. Therefore, we have searched infectious disease names and outbreak locations in eKG using the BioPortal Annotator API (https://data.bioontology.org/documentation#nav_annotator) and the GeoNames Search Webservice (https://www.geonames.org/export/geonames-search.html), respectively, and the discovered entities have been stored in eKG. Specifically, infectious disease names identified in the epidemiological reports have been linked to biomedical ontology entities using the `skos:related` axiom, while the `obo:RO_0001025` property, a.k.a. `located in`, has been adopted for linking the geo-locations to the equivalent GeoNames entities.

The outcome of the entity linking process is twofold. First, it allows eKG to be enriched with additional information from external domain ontologies and GeoNames. For example, if a disease outbreak in eKG is linked to a biomedical entity in the BioPortal, additional information about the disease outbreak such as its symptoms or its transmission patterns could be quickly accessed and analysed. Second, it provides a mechanism for connecting eKG to other knowledge graphs and datasets that are using standard biomedical terminologies. This is important for enabling cross-domain knowledge discovery, integration, and maximising interoperability.

The inclusion of links to external domain ontologies in the eKG entity metadata provides additional contextual information and connections to external knowledge sources, allowing for more comprehensive and diverse data analysis. This linkage allows researchers and public health professionals to draw on a wealth of existing ontological information to better understand the outbreak's characteristics and study effective response strategies. The inclusion of external links also facilitates the integration of eKG with other knowledge graphs and databases, enabling cross-disciplinary research and collaboration.

An in-depth explanation of the structure of eKG and each associated data record is provided in the "Data Records" section, which also offers an overview of the data files and formats. The disease outbreak events have been extracted from the DONs reports by the LLMs ensemble and mapped as individuals of eKG, leveraging the computational power of the JRC Big Data Analytics Platform[76]. The epidemiological knowledge graph is made available[77] in structured RDF/XML and Turtle (https://www.w3.org/TR/turtle/) formats, with raw extractions also given in CSV, within the Joint Research Centre Data Catalogue (https://data.jrc.ec.europa.eu/) at the following permanent location: http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-f99d1c0f6601, used also as reference namespace for the eKG individuals. The knowledge graph was also released within the European

Data portal[78], the official data repository for European data, at the following permanent location: https://data.europa.eu/data/datasets/89056048-7f5d-4d7c-96ad-f99d1c0f6601.

eKG Services & Interfaces. Using the described pipeline, we generated the ontologically-grounded knowledge graph of epidemiological information, comprising roughly of 46K distinct, non-reified (i.e. that are not given additional metadata or context, meaning they are expressed in a straightforward `<subject, predicate, object>` format without any further elaboration or annotation[21]) triples, representing around 2.9K unique outbreak events via a total of 3.6K generalized axioms, as of the 28th February 2025. To access the data programmatically, we set up a *Virtuoso SPARQLendpoint*, available at: https://api-vast.jrc.service.ec.europa.eu/sparql/, where eKG can be queried and analytical information on target disease outbreaks, attributes, and relations can be retrieved in structured data format using the SPARQL query language[19]. On top of eKG we are currently developing a number of intuitive and user-friendly services and interfaces at European Commission premises, including content negotiation, visualization and navigation of data, improved exploitation of the available information and knowledge discovery. These semantic services are applicable and valuable to any kind of knowledge graph based on RDF/OWL technology, enhancing their utility and accessibility across various domains. The central point of access (Fig. 3) is a web interface where it is possible to have an immediate view of the main epidemiological information associated to a disease outbreak, along with quick reference pointers to other semantic applications (see also the "Usage Notes" section), namely: the *Virtuoso Faceted Browser*, available at https://api-vast.jrc.service.ec.europa.eu/fct/, a general purpose RDF data query facility service for faceted browsing over entity relationship types, allowing users to explore the outbreak events in a structured and intuitive way (e.g., see the left snapshot of Fig. 3); *LodView*[79], a web application compliant to World Wide Web Consortium (W3C)[80] standards for IRI dereferentiation which provides a representation of our RDF disease outbreak resource through a custom intuitive HTML page (http://lodview.it/, e.g. see the bottom snapshot of Fig. 3); *LodLive*[81], a navigator of RDF resources based on a graph layout (http://lodlive.it/, e.g. see the top-right snapshot of Fig. 3).

Data Records

The extracted epidemiological information from DONs is available in CSV format, while both the DONs and the knowledge graph are available in RDF/XML and Turtle formats within the Joint Research Centre Data Catalogue[77]. Data are freely available to the public and can be downloaded without any login details by selecting the appropriate URL in one of the repositories above. The raw CSV file available in the repository[77] contains the extracted details of the occurring event from each report. In particular, it contains the following information:

- "fileid" column, presents the ID of the report, which is taken from the slug of the URL of the DON item;
- "virus_extracted" column, denotes the extracted name of the disease outbreak extracted from the current DON item;
- "country_extracted" column, depicts the involved countries as extracted from the report linked to the specific disease outbreak;
- "date_extracted" column, presents the date when the disease outbreak occurred, in the format YYYY/MM/DD;
- "date_cases_Imputed" column, shows the date reported in the fileid of the report, if present, and it can be related to the publication date of the report (YYYY/MM/DD format), although not formally defined in DONs. It is important to note that the date in the report may vary significantly from the actual date of the disease outbreak. Therefore, this information could be useful for analysis only when the "date_extracted" is missing, which means it was not possible to automatically extract the actual outbreak date from the main text of the report;
- "cases_extracted" column, shows the extracted number of confirmed cases deriving from the specific disease outbreak;
- "deaths_extracted" column, presents the related event mortality.

In our data engineering work to model the knowledge graph of epidemiological information, we adhered to standard style guidelines for identifying and describing linked data resources[82]. Specifically, within eKG, data names are formatted in lowercase, with any spaces replaced by dashes, adhering to established conventions for naming and labeling knowledge graphs[16,82]. Conversely, class names in the ontology definitions are formatted in uppercase, in line with common style guidelines for such terminologies[16,82]. In this context, data names refer to the specific identifiers assigned to data elements within the ABox of the knowledge graph, while class names in the ontology definitions are part of its TBox, defining the structure and constraints of eKG[14].

The main classes of the knowledge graph are mapped to RDF/OWL, as illustrated in Fig. 4. A disease outbreak event from DONs is an instance (`rdf:type`) of the `IDO:surveillance_process` class, which is a specialization of a generic `OBI_0000011:planned_process`. A disease outbreak is associated with an `IDO_0000436:infectious_disease`, which has a specific `infectious_disease_name` that corresponds to the name of the disease extracted from the DONs text, occurring in a `GEO_000000372:geographical_region` on a `date_of _infection`, i.e., a `dcterms:date`[83]. Please note that when a date is reported in the title of a DONs, such as in the "Nipah virus - India" DON of the 31st May 2018 as an example, an additional `imputed_date_of_infection` is also associated to the disease outbreak event; however, this information should be taken with the due care, given the reporting date might be delayed with respect to the actual occurring date of the disease outbreak. In the "Nipah virus - India" DON example, indeed, the reporting `imputed_date_of_infection` would be the "2018-05-31", but the actual `date_of_infection` reported in the DON is the "2018-05-19". The `imputed_date_of_infection`

Top 10 Disease Outbreak Names	Top 10 Countries	Top 10 Disease Outbreak Names - Country pairs
Avian influenza (570)	China (243)	**Saudi Arabia - MERSCoV** (201)
MERSCoV (296)	Democratic Republic of Congo (226)	China - Avian influenza (172)
Ebola (211)	Saudi Arabia (225)	**Democratic Republic of Congo - Ebola** (134)
Cholera (146)	Egypt (155)	The Netherlands - Avian influenza (108)
Yellow Fever (146)	The Netherlands (142)	Egypt - Avian influenza (101)
SARS (78)	United Republic of Tanzania (89)	**China-SARS** (51)
Hemorrhagic fever (68)	Viet Nam (66)	Viet Nam - Avian influenza (41)
Dengue Fever (56)	Angola (60)	Democratic Republic of Congo - Hemorrhagic fever (31)
Meningococcal (Group C) (54)	Cameroon (51)	**Guinea - Ebola**(30)
H5N1 (51)	Guinea Bissau (49)	Cameroon - Yellow Fever (30)

Table 1. The top reported disease outbreak names, countries, and disease outbreak name-country pairs by total number of extracted events (number given in parentheses). Pairs of disease outbreak name-country in **bold** are further illustrated in Fig. 5.

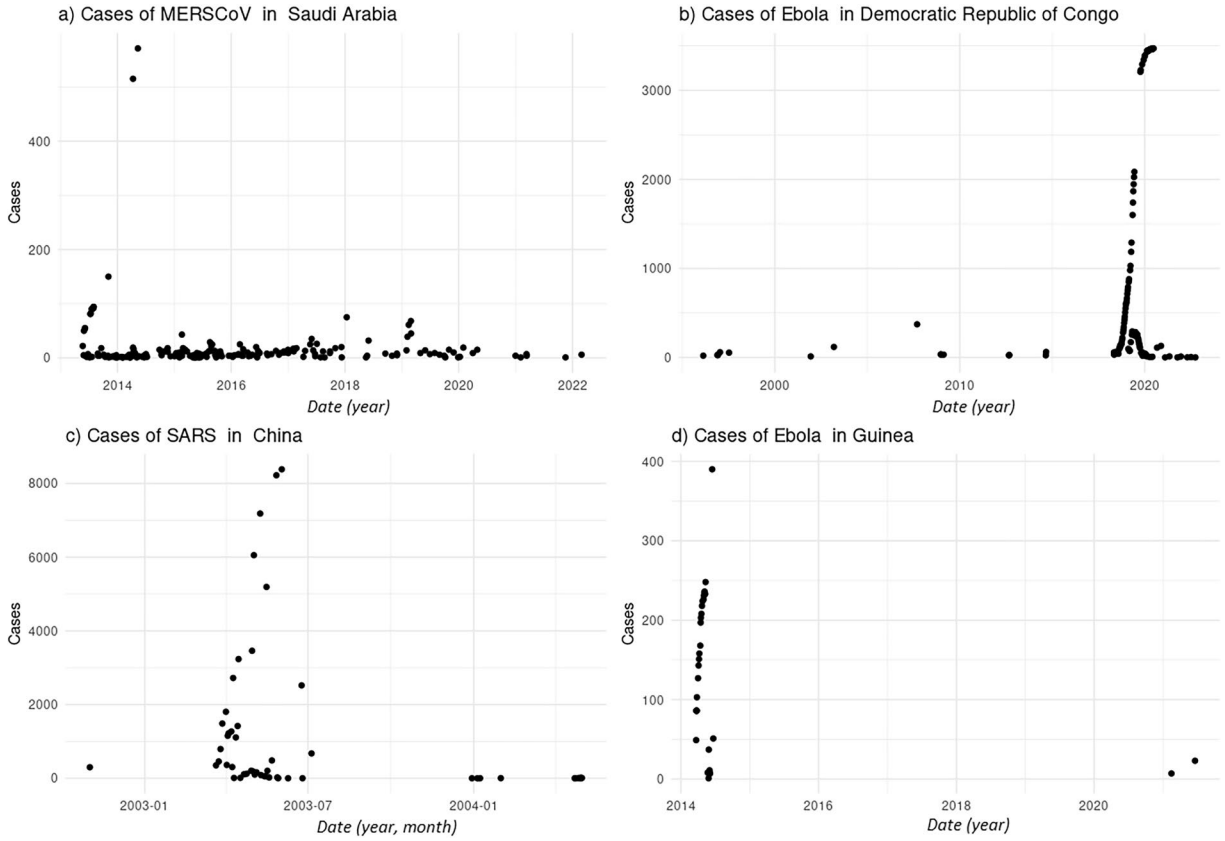

Fig. 5 Number of extracted cases vs. time for 4 case studies: MERS-Cov in Saudi Arabia (Panel a), Ebola cases in the Democratic Republic of Congo (Panel b), SARS cases in China (Panel c), and Ebola cases in Guinea (Panel d). These four case studies correspond to the Disease Outbreak Names - Country pairs reported in bold in Table 1.

could be useful for analysis only when it was not possible to automatically extract from the main text of the report an actual `date_of_infection`. A disease outbreak event is then associated with a portion of `infected_population` expressed in terms of `IDO_0000511:infected _population_cases` extracted from the DON report, which can ultimately lead to an `IDO_0000489:infectious _disease_mortality`, i.e., the extracted number of deaths from the disease outbreak report. It is worth noting that the `skos:related` triples serve to connect the "Nipah virus" disease outbreak, identified in the report, with corresponding outbreak names from various open-source vocabularies on the web through entity linking. Additionally, the triple having `obo:RO_0001025` property (which denotes the relationship `located in`) associates the location "India" with its equivalent GeoNames identifier, i.e. `geonames:1269750`, also facilitated by entity linking.

	Precision	Recall	F_1 score
GPT-3.5-Turbo-16k	0.750	0.961	0.842
GPT-4-32k	0.756	0.945	0.840
GPT-4-FewShot	0.770	0.914	0.836
Pythia-12B	0.712	0.773	0.742
MPT-30B-Chat	0.740	0.891	0.809
Meta-Llama-2-70B-Chat	0.757	0.898	0.821
Meta-Llama-3-70B-Instruct	0.762	0.926	0.834
Mistral-7B-OpenOrca	0.755	0.891	0.817
Zephyr-7B-Alpha	0.752	0.875	0.809
Zephyr-7B-Beta	0.756	0.882	0.816
Ensemble	**0.794**	**0.988**	**0.851**

Table 2. LLMs comparison for the information extraction task of the disease outbreak name over IDB.

Consider the following vocabulary prefixes, in RDF/XML:

```
<rdf:RDF
  xmlns="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-96ad-
  99d1c0f6601/"
  xml:base="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-4d7c-
  96ad-f99d1c0f6601/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xml="http://www.w3.org/XML/1998/namespace"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:obo="http://purl.obolibrary.org/obo/"
 >
```

The RDF/XML triples of this example are encapsulated within the following OWL statement:

```
<rdf:Description rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-
7f5d-4d7c-96ad-f99d1c0f6601/don-record2740">


  <rdf:type rdf:resource="http://purl.obolibrary.org/obo/VSMO_" />
  <dcterms:identifier>don-record2740</dcterms:identifierl><dcterms:descrip-
  tion>Outbreak of Nipah Virus occurred in India on 2018-05-19, with con-
  firmed 15 cases and 13 deaths.</dcterms:description><rdfs:label>31-may
  -2018-nipah-virus-india-en</rdfs:label><virus_extracted>Nipah Virus</
  virus_extracted>
  <skos:related rdf:resource="http://purl.bioontology.org/ontology/MESH/
  D045405" />
  <skos:related rdf:resource="http://ncicb.nci.nih.gov/xml/owl/EVS/
  Thesaurus.owl#C112359" />
  <skos:related rdf:resource="http://purl.bioontology.org/ontology/SNOMEDCT
  /115511007"/>
  <skos:related rdf:resource="http://purl.obolibrary.org/obo/NCBITaxon_
  3052225" />
  <skos:related rdf:resource="http://sbmi.uth.tmc.edu/ontology/ochv#C0
  751673/>
  <country_extracted>India</country_extracted>
  <obo:RO_0001025 rdf:resource="https://sws.geonames.org/1269750/"/>
  <date_extracted rdf:datatype="http://www.w3.org/2001/XMLSchema#
  date">2018-05-19</date_extracted>
  <date_cases_Imputed rdf:datatype="http://www.w3.org/2001/XMLSchema#
  date">2018-05-31</date_cases_Imputed>
  <cases_extracted>15</cases_extracted><deaths_extracted>13</deaths_
  extracted></owl:Class>
```

	Precision	Recall	F_1 score
GPT-3.5-Turbo-16k	0.981	0.910	0.944
GPT-4-32k	0.981	0.928	0.954
GPT-4-FewShot	0.981	0.928	0.954
Pythia-12B	0.962	0.452	0.615
MPT-30B-Chat	0.979	0.855	0.913
Meta-Llama-2-70B-Chat	0.980	0.898	0.937
Meta-Llama-3-70B-Instruct	0.984	0.926	0.956
Mistral-7B-OpenOrca	0.986	0.861	0.920
Zephyr-7B-Alpha	0.981	0.910	0.944
Zephyr-7B-Beta	0.982	0.914	0.950
Ensemble	**0.988**	**0.931**	**0.962**

Table 3. LLMs comparison for the information extraction task of the country name over IDB.

	Precision	Recall	F_1 score
GPT-3.5-Turbo-16k	0.699	0.455	0.551
GPT-4-32k	0.673	0.589	0.629
GPT-4-FewShot	0.733	0.589	0.653
Pythia-12B	0.365	0.277	0.315
MPT-30B-Chat	0.619	0.464	0.531
Meta-Llama-2-70B-Chat	0.561	0.536	0.548
Meta-Llama-3-70B-Instruct	0.722	0.590	0.652
Mistral-7B-OpenOrca	0.67	0.527	0.590
Zephyr-7B-Alpha	0.615	0.571	0.593
Zephyr-7B-Beta	0.650	0.585	0.612
Ensemble	**0.788**	**0.591**	**0.658**

Table 4. LLMs comparison for the information extraction task of the number of confirmed cases over IDB.

	Precision	Recall	F_1 score
GPT-3.5-Turbo-16k	0.902	0.639	0.748
GPT-4-32k	0.913	0.734	0.814
GPT-4-FewShot	0.904	0.658	0.762
Pythia-12B	0.735	0.158	0.260
MPT-30B-Chat	0.814	0.304	0.442
Meta-Llama-2-70B-Chat	0.916	0.759	0.830
Meta-Llama-3-70B-Instruct	0.922	0.807	0.862
Mistral-7B-OpenOrca	0.908	0.690	0.784
Zephyr-7B-Alpha	0.920	0.804	0.858
Zephyr-7B-Beta	0.921	0.810	0.861
Ensemble	**0.928**	**0.813**	**0.869**

Table 5. LLMs comparison for the information extraction task of the disease outbreak date over IDB.

where the raw extractions metadata are linked the ontology classes of eKG as shown below:

```
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-
 4d7c-96ad-f99d1c0f6601/virus_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/IDO_
 0000436"/></owl:Class>
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-
 4d7c-96ad-f99d1c0f6601/country_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/GEO_
 000000372"/></owl:Class>
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-
 4d7c-96ad-f99d1c0f6601/date_extracted">
<rdfs:subClassOf rdf:resource="http://purl.org/dc/terms/date"/>
</owl:Class>
```

```
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-
 4d7c-96ad-f99d1c0f6601/date_cases_Imputed">
<rdfs:subClassOf rdf:resource="http://purl.org/dc/terms/date"/>
</owl:Class>
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-
 4d7c-96ad-f99d1c0f6601/cases_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/IDO_
 0000511"/></owl:Class>
<owl:Class rdf:about="http://data.jrc.ec.europa.eu/dataset/89056048-7f5d-
4d7c-96ad-f99d1c0f6601/deaths_extracted">
<rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/IDO_
0000489"/></owl:Class>
```

We enable direct access to the extracted knowledge graph data via custom SPARQL queries through a dedicated REST API SPARQL endpoint.

Data Statistics. After removing 353 entries with missing country and disease, the dataset consists of 2556 entries, listing 144 unique pathogens/diseases and 207 unique countries. Table 1 shows the top 10 disease outbreak names, countries, and disease outbreak names - country pairs by total number of extracted events. The events detected contain a combination of human diseases and zoonoses, with worldwide coverage. Figure 5 presents four case studies as a time series of the number of cases over time. Figure 5, Panel a, shows the numbers of MERS-Cov cases reconstructed by the LLM in Saudi Arabia, with a consistent reconstruction of the MERS-Cov cases experienced in the area (https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON516). Panels b) and d) of Fig. 5 show the number of Ebola cases reconstructed by the LLM in the Democratic Republic of the Congo (DRC) and Guinea, respectively, illustrating the Kivu Ebola Epidemic of 2018-2020 in DRC (https://www.who.int/emergencies/situations/Ebola-2019-drc-) and the West Africa outbreak of 2014-2016 (https://www.who.int/emergencies/situations/ebola-outbreak-2014-2016-West-Africa). Panel c) of Fig. 5 shows instead the reconstructed number of SARS cases in China, with a qualitative reconstruction of the SARS outbreak of 2002–2004 (https://en.wikipedia.org/w/index.php?title=2002%E2%80%932004_SARS_outbreak). Overall, the dataset shows potential for reconstructing trends for some epidemics, particularly those involving diseases included in the International Health Regulations, as these are consistently reported and updated in the DONs. For an in-depth quantitative comparison, see the next "Technical Validation" section.

Technical Validation

In this section, we present a formal evaluation of the quality of the extracted epidemiological data within eKG. While the generation process of the knowledge graph is entirely automated, it is important to assess the accuracy of the extraction methodology. This evaluation will serve as a support to the dissemination of eKG as a public resource. Building on the work in the literature[10], in this section we present a comparison of the performance of different LLMs for infectious disease information extraction in order to validate the adopted ensemble approach for producing eKG. The evaluation was conducted on a subset of the Incident Database (IDB), which is a repository created specifically for event-based surveillance[84]. The IDB is structured to house a comprehensive collection of data related to epidemic events, making it an important resource for researchers and public health professionals. The database includes reports from various surveillance systems, including samples from WHO DONs, among others, that were meticulously annotated by domain experts. These annotations provide crucial insights, allowing researchers to easily identify the key elements of the reports. Prior to integration into the IDB, the data is subjected to a thorough cleaning and standardization process. This procedure guarantees that every entry in the database adheres to a uniform format, simplifying data analysis and interpretation. Furthermore, data initially presented in foreign languages is translated to make it accessible to a wider audience. To evaluate the performance of the epidemiological information extraction models under investigation, a reference subset of the IDB, consisting of 171 samples for which we were able to reconstruct the source url and locate the main text of the report, was considered. This subset serves as a gold standard for benchmarking the performance of the LLMs.

We differentiate the comparison for each of the epidemiological IE tasks, namely the extraction of the name of the disease, the confirmed case count, the country where the cases were reported, and the date of the case count. It was not possible to evaluate the task of extracting the number of deaths as this information was not annotated in the IDB. Nonetheless, given the similarity of this task to the extraction of the number of cases, its performance is likely to be comparable. However, it's important to note that the variability in terms used to refer to deaths – such as fatalities, casualties, deceased, and passed away – may impact the comparability with extracting the number of cases.

Each IE task was assessed as a binary classification problem. In this context, the negative class signifies that no information ("None") is associated with the IDB data sample, whereas the positive class indicates that the IDB is tagged with specific epidemiological information. The models were evaluated on their capability to identify these positive and negative classes using the commonly used metrics of *Precision*, *Recall*, and F_1 *score* for each class[85]:

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

$$F_1 = \frac{TP}{TP + (FP + FN)/2} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}, \tag{3}$$

where *TP*, *FP*, *FN*, and *TN* denote the counts of True Positives, False Positives, False Negatives, and True Negatives, respectively[85]. In simple terms, *Precision* (or *Positive Predictive Value*) refers to the percentage of positively classified instances that are correctly identified. *Recall* (or *Sensitivity*) is the fraction of positive class instances that the classifier accurately identifies. The F_1*score*, which is the harmonic mean of precision and recall, is favoured over traditional accuracy in imbalanced problem settings because it places greater emphasis on lower values, unlike the regular mean which treats all values equally. A high F_1 score for the positive class is achieved only when both its precision and recall are large.

For the computational analysis, we considered the LLMs used in *Ensemble*, namely *Mistral-7B-OpenOrca*, *Zephyr-7B-Beta*, and *Meta-Llama-3-70B-Instruct*, as well as the other open-source LLMs versions adopted in the literature[10], namely *Pythia-12B*[86], *MPT-30B-Chat* (https://huggingface.co/mosaicml/MPT-30B-Chat), and *Meta-Llama-2-70B-Chat*. We further considered in the analysis the commercial OpenAI's GPT models family (https://platform.openai.com/docs/models), namely *GPT-3.5-Turbo-16k* and *GPT-4-32k*, along with a *GPT-4-FewShot* variant of the *GPT-4-32k* model, adapted to handle specific epidemic information extraction tasks. This adaptation is achieved by providing the model with a limited number of examples, leveraging in-context learning (ICL)[9] to enable the model to learn from these examples and generate accurate responses. Unlike supervised approaches, which involve a training phase that uses backward gradients to adjust model parameters, ICL skips parameter adjustments and directly performs predictions using pretrained LLMs. The anticipation is that the model will discern the hidden pattern in the provided examples and produce a more precise output[87]. The *GPT-4-FewShot* variant, similarly, utilizes in-context learning to specialize in epidemiological information extraction tasks, allowing it to learn from a few contextual examples (three-shot, in our case) and generate more accurate responses. The ICL strategy was exclusively applied to *GPT-4-32k* because of its extensive context capacity, which was not feasible with the other LLMs studied, as they have significantly lower context capabilities.

For our LLMs experiments we leveraged the GPT@JRC platform (https://commission.europa.eu/news/commission-launches-new-general-purpose-ai-tool-gptec-2024-10-22_en), allowing JRC personnel to investigate the possible applications of AI pre-trained LLMs. This initiative is a key element of a wider investigation into how the European Commission can utilize this emerging technology. Serving as a main hub, it provides secure access to a variety of AI models. The GPT@JRC platform is physically located in the JRC datacentre and supports both open-source AI models, which are deployed on-site at the JRC Big Data Analytics Platform[76], and commercial OpenAI GPT models, which run in the European Cloud under a Commission contract that includes an option to opt-out of third-party prompt analysis.

Tables 2–5 report our computational results on the comparison of the LLMs for the extraction of, respectively, the disease outbreak names, the countries where the outbreak occurred, the related event date, and the number of confirmed cases associated to the specific disease outbreak. Our analysis reveals that the *Pythia-12B* and *MPT-30B-Chat* models exhibit lower performance compared to their counterparts. These models, while still valuable, may not be the optimal choice for applications requiring the highest level of accuracy in epidemiological IE tasks.

Commercial LLMs, particularly *GPT-4-32k*, demonstrated superior performance, aligning with expectations given their advanced architecture and extensive training datasets. *GPT-3.5-Turbo-16k* also shows a significant performance, indicating its potential utility in epidemiological IE tasks. The application of in-context learning, specifically the *GPT-4-32k-FewShot* variant, showed that providing the model with a few examples can enhance its performance, although this improvement is not consistent across all tasks.

Despite the high performance of OpenAI's GPT models, their implementation is associated with significant costs and practical limitations, such as usage restrictions, which may hinder their widespread deployment on large datasets like the full DONs.

In the realm of open-source models, we observe notable advancements in performance with newer iterations. The *Meta-Llama-3-70B-Instruct* model outperforms its predecessor, *Meta-Llama-2-70B-Chat*, and similarly, *Zephyr-7B-Beta* shows improvements over *Zephyr-7B-Alpha*, as we were expecting. These enhancements underscore the rapid progress within the field of open-source LLMs.

Among the open-source LLMs, *Meta-Llama-3-70B-Instruct*, *Mistral-7B-OpenOrca*, and *Zephyr-7B-Beta* stand out for their high performance, rivaling and occasionally surpassing that of the commercial GPT models. In our computational experience the *Meta-Llama-3-70B-Instruct* model exhibited slower inference times due to its large number of parameters, which may impact its practicality in time-sensitive applications.

The *Ensemble* approach, which combines the strengths of *Mistral-7B-OpenOrca*, *Meta-Llama-3-70B-Instruct*, and *Zephyr-7B-Beta*, emerges as the most effective strategy (see results in bold in the tables). This method not only achieves the highest F_1 scores across all four tasks but also mitigates the limitations of individual models. The ensemble's robust performance across diverse IE tasks suggests its suitability for full-scale deployment on the entire DONs dataset, providing a reliable and efficient solution for extracting structured epidemiological information, and supporting the technical quality of the extracted eKG dataset. The *Ensemble* approach leverages the collective capabilities of high-performing open-source LLMs to deliver an optimal balance of accuracy, speed, and adaptability. The superior performance of this method, as evidenced by the results in Tables 2–5, supports the reliability of the produced dataset and its large potential for epidemic outbreak modeling and surveillance systems.

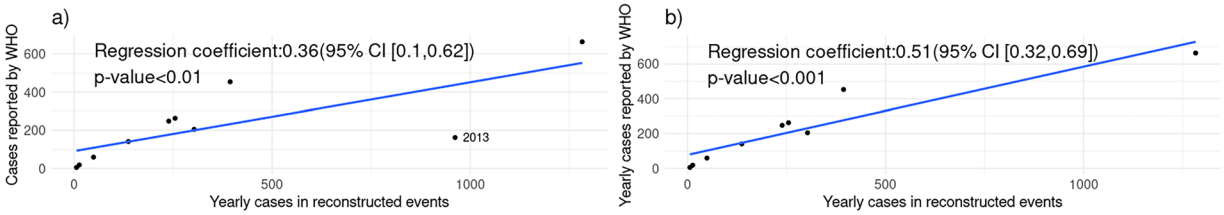

Fig. 6 WHO yearly number of cases vs the number of cases from the reconstructed events, together with regression line: including year 2013 (Panel a) and without year 2013 (Panel b). Since 2013 was the epidemic's onset, the inclusion of potential and confirmed cases by WHO DONS could lead to a lower correlation coefficient due to overestimated reconstructed cases. Correlation and p-values are reported in the top right corner of each panel.

From a qualitative perspective, we compared the yearly number of MERS-Cov cases in Saudi Arabia from the reconstructed events with the number of confirmed cases reported by the WHO. Figure 6 shows the scatter plots of WHO yearly number of cases vs the number of cases from the reconstructed events, together with the best-fit regression line, the 95% CI of the regression coefficient and the corresponding p-value. Panel a shows the results for all data, while Panel b depicts the results obtained by excluding year 2013. As it can be observed, the regression coefficient between the reconstructed events and the WHO data is significant (Panel a: 95% CI of regression coefficient [0.10, 0.62], p-value < 0.01) and increases considerably (Panel b: 95% CI of regression coefficient [0.30, 0.69], p-value < 0.001) when data from year 2013 is not considered. Considering 2013 was the onset of the epidemic, it is plausible that WHO DONS might have included potential cases along with confirmed ones. This could result in an overestimation of the reconstructed cases compared to the confirmed ones, resulting in a lower regression coefficient if data from year 2013 are taken into consideration.

Usage Notes

Our open-access JRC Data Catalogue[77] contains the complete eKG knowledge graph in both RDF/XML and Turtle formats, along with the raw extractions from which the knowledge graph has been derived, in CSV format. We enable the interested community to access and use the produced data and ontology under Creative Commons Attribution 4.0 International license (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/). All the other ontologies mentioned in the paper that have been used to construct and map the eKG knowledge graph are also available under CC BY 4.0 license under their sources, and, respectively: Infectious Disease Ontology (IDO), at https://www.ebi.ac.uk/ols4/ontologies/ido; GeoNames, at https://www.geonames.org/; Basic Formal Ontology (BFO), at https://basic-formal-ontology.org/; and OBO Foundry, at https://obofoundry.org/. These ontologies provide structured entities and terms that are relevant to the infectious disease domain, geographical entities, and formal ontological structures, which are incorporated into the eKG.

To perform queries to eKG, a user can download and import the entire knowledge graph in one of the available formats into an OpenLink Virtuoso triplestore (https://virtuoso.openlinksw.com/, minimum version 7.0) instanced on a local machine. After downloading and installing Virtuoso, users should access the server conductor interface through a web browser. They will navigate to the quad store upload section for linked data. By creating a new graph, users can configure it by specifying a name and uploading the entire RDF/XML or Turtle file to the newly created knowledge graph. Users can verify if the dataset has been successfully populated by executing queries using the Virtuoso SPARQL query interface. For convenience, in addition to the repository data, eKG is also available for users to be interrogated in an OpenLink Virtuoso instance that we have instantiated at https://api-vast.jrc.service.ec.europa.eu/sparql/, and explored in a structured and intuitive way via the related Virtuoso Faceted Browser (https://api-vast.jrc.service.ec.europa.eu/fct/). Other visualization services can be leveraged at this purpose, such as *LodView*[79], available at http://lodview.it/, which implements content negotiation of eKG resources and allows to download the disease outbreaks in various needed formats, such as, e.g., *xml*, *ntriples*, *turtle*, and *ld+json*, and *LodLive*[81], available at http://lodlive.it/, which enables the navigation of eKG resources via an effective graph representation, starting from a specific disease outbreak event and expanding automatically towards its relationships.

The approach we utilized in constructing eKG allowed us to develop a high-quality knowledge graph encompassing trustworthy epidemic events. These events have been validated through experimental methods and strongly endorsed by data providers and collaborating EU institutions. Furthermore, we have meticulously verified the meaning of these events to ensure a consistent representation of domain knowledge.

Code availability

The codebase and documentation of the pipeline used to extract epidemiological information from WHO DONs using the *Ensemble* of LLMs, and to generate the eKG, can be found at: https://github.com/jrcf7/epidemicIE-eKG-dons. The code was written in Python 3.10.11, and we used packages like Pandas (https://pandas.pydata.org/) for handling the raw extraction files (e.g. CSV), Numpy (https://numpy.org/) for data preprocessing and feature extraction, Scikit-learn (https://scikit-learn.org/) for common machine learning routines, WordNet (https://wordnet.princeton.edu/) as a lexical database for the English language, available via NLTK (https://www.nltk.org/), assisting in word sense disambiguation and synonymy checking, Sentence Transformers SBERT (Sentence-BERT: https://sbert.net/) to derive semantically meaningful sentence embeddings for semantic similarity at the sentence level, the PyTorch (https://pytorch.org/) framework for handling the LLMs workflow, and LangChain (https://www.langchain.com/) for data processing and

integration with the LLMs through modular components and pre-built libraries. We used the European Commission GPT@JRC platform to have secure access to the various pre-trained LLMs employed in this study. Finally, the *Protégé* editor (http://protege.stanford.edu) was employed to manage, adjust and maintain the knowledge graph structure.

Received: 15 October 2024; Accepted: 27 May 2025;
Published online: 10 June 2025

References

1. Salathé, M. *et al*. Digital epidemiology. *PLoS Computational Biology* **8**, e1002616, https://doi.org/10.1371/journal.pcbi.1002616 (2012).
2. Brownstein, J. S., Rader, B., Astley, C. M. & Tian, H. Advances in Artificial Intelligence for Infectious-Disease Surveillance. *New England Journal of Medicine* **388**, 1597–1607, https://doi.org/10.1056/NEJMra2119215 (2023).
3. Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D. & Weinstein, R. A. Using internet searches for influenza surveillance. *Clinical infectious diseases* **47**, 1443–1448, https://doi.org/10.1086/593098 (2008).
4. Warsame, A., Murray, J., Gimma, A. & Checchi, F. The practice of evaluating epidemic response in humanitarian and low-income settings: a systematic review. *BMC Medicine* **18**, https://doi.org/10.1186/s12916-020-01767-8 (2020).
5. Oppenheim, B. *et al*. Assessing global preparedness for the next pandemic: Development and application of an Epidemic Preparedness Index. *BMJ Global Health* **4**, https://doi.org/10.1136/bmjgh-2018-001157 (2019).
6. Mondor, L. *et al*. Timeliness of nongovernmental versus governmental global outbreak communications. *Emerging Infectious Diseases* **18**, 1184–1187, https://doi.org/10.3201/eid1807.120249 (2012).
7. Lugo-Robles, R., Garges, E., Olsen, C. & Brett-Major, D. Identifying nontraditional epidemic disease risk factors associated with major health events from world health organization and world bank open data. *American Journal of Tropical Medicine and Hygiene* **105**, 896–902, https://doi.org/10.4269/ajtmh.20-1318 (2021).
8. Vaswani, A. *et al*. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5999–6009, https://doi.org/10.5555/3295222.3295349 (2017).
9. Brown, T. B. *et al*. Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901, https://doi.org/10.5555/3495724.3495883 (2020).
10. Consoli, S. *et al*. Epidemic Information Extraction for Event-Based Surveillance Using Large Language Models. In *Proceedings of Ninth International Congress on Information and Communication Technology (ICICT 2024)*, vol. 1011 LNNS, 241–252, https://doi.org/10.1007/978-981-97-4581-4_17 (Lecture Notes in Networks and Systems, 2024).
11. Heath, T. & Bizer, C. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology* **1**, 1–121, https://doi.org/10.2200/S00334ED1V01Y201102WBE001 (2011).
12. Auer, S. *et al*. Towards a Knowledge Graph for Science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS '18, https://doi.org/10.1145/3227609.3227689 (Association for Computing Machinery, New York, NY, USA, 2018).
13. Peng, C., Xia, F., Naseriparsa, M. & Osborne, F. Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review* **56**, https://doi.org/10.1007/s10462-023-10465-9 (2023).
14. Ji, S., Pan, S., Cambria, E., Marttinen, P. & Yu, P. S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems* **33**, 494–514, https://doi.org/10.1109/TNNLS.2021.3070843 (2022).
15. Tiwari, S., Ortíz-Rodriguez, F., Abbés, S. B., Usip, P. U. & Hantach, R. *Semantic AI in Knowledge Graphs* (Taylor & Francis, Boca Raton, US, https://doi.org/10.1201/9781003313267 2023).
16. Hogan, A. *et al*. Knowledge graphs. *ACM Computing Surveys* **54**, https://doi.org/10.1145/3447772 (2021).
17. Baader, F., Horrocks, I., Lutz, C. & Sattler, U.*An Introduction to Description Logic* (Cambridge University Press, 2017).
18. Antoniou, G. & van Harmelen, F. Web Ontology Language: OWL. In *Handbook on Ontologies*, 67–92, https://doi.org/10.1007/978-3-540-24750-0_4 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004).
19. Pérez, J., Arenas, M. & Gutierrez, C. Semantics and complexity of SPARQL. *ACM Transactions on Database Systems* **34**, 1–45, https://doi.org/10.1145/1567274.1567278 (2009).
20. Alocci, D. *et al*. Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. *PLOS ONE* **10**, 1–17, https://doi.org/10.1371/journal.pone.0144578 (2015).
21. Orlandi, F., Graux, D. & O'sullivan, D. Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF*. In *Proceedings-2021 IEEE 15th International Conference on Semantic Computing, ICSC 2021*, 233–240, https://doi.org/10.1109/ICSC50631.2021.00049 (2021).
22. Maynard, D., Bontcheva, K. & Augenstein, I. Natural language processing for the Semantic Web. *Synthesis Lectures on the Semantic Web: Theory and Technology* **6**, 1–196, https://doi.org/10.1007/978-3-031-79474-2 (2017).
23. Zhou, Z.-H. *Ensemble methods: Foundations and algorithms* (Chapman & Hall/CRC, https://doi.org/10.1201/b12207 2012).
24. Halevy, A. Information integration. In Liu, L. & Özsu, M. T. (eds.) *Encyclopedia of Database Systems*, 1490–1496, https://doi.org/10.1007/978-0-387-39940-9_1069 (Springer US, Boston, MA, 2009).
25. Carmen Legaz-García, M. D., Miñarro-Giménez, J. A., Menárguez-Tortosa, M. & Fernández-Breis, J. T. Generation of open biomedical datasets through ontology-driven transformation and integration processes. *Journal of Biomedical Semantics* **7**, https://doi.org/10.1186/s13326-016-0075-z (2016).
26. Mesiti, M. *et al*. XML-based approaches for the integration of heterogeneous bio-molecular data. *BMC Bioinformatics* **10**, S7, https://doi.org/10.1186/1471-2105-10-S12-S7 (2009).
27. Bonfitto, S., Casiraghi, E. & Mesiti, M. Table understanding approaches for extracting knowledge from heterogeneous tables. *WIREs Data Mining and Knowledge Discovery* **11**, e1407, https://doi.org/10.1002/widm.1407 (2021).
28. Parundekar, R., Knoblock, C. A. & Ambite, J. L. Linking and building ontologies of linked data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6496 LNCS**, 598–614, https://doi.org/10.1007/978-3-642-17746-0_38 (2010).
29. Poggi, A. *et al*. Linking data to ontologies. In Spaccapietra, S. (ed.) *Journal on Data Semantics X*, 133–173, https://doi.org/10.1007/978-3-540-77688-8_5 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
30. Mahmud, S. M. H. *et al*. Publishing csv data as linked data on the web. *Lecture Notes in Electrical Engineering* **605**, 805–817, https://doi.org/10.1007/978-3-030-30577-2_72 (2020).
31. Lefrançois, M., Zimmermann, A. & Bakerally, N. A SPARQL extension for generating RDF from heterogeneous formats. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10249 LNCS**, 35-50, https://doi.org/10.1007/978-3-319-58068-5_3 (2017).
32. Rodríguez-Muro, M. & Rezk, M. Efficient SPARQL-to-SQL with R2RML mappings. *Journal of Web Semantics* **33**, 141–169, https://doi.org/10.1016/j.websem.2015.03.001 (2015).
33. Iglesias-Molina, A. *et al*. The RML Ontology: A Community-Driven Modular Redesign After a Decade of Experience in Mapping Heterogeneous Data to RDF. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **14266 LNCS**, 152–175, https://doi.org/10.1016/10.1007/978-3-031-47243-5_9 (2023).

34. García-González, H., Boneva, I., Staworko, S., Labra-Gayo, J. E. & Lovelle, J. M. C. ShExML: improving the usability of heterogeneous data mapping languages for first-time users. *PeerJ Computer Science* **6**, https://doi.org/10.7717/PEERJ-CS.318 (2020).
35. Heyvaert, P., De Meester, B., Dimou, A. & Verborgh, R. Declarative rules for linked data generation at your fingertips! *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11155 LNCS**, 213–217, https://doi.org/10.1007/978-3-319-98192-5_40 (2018).
36. Zhang, S. *et al.* A Graph-based Approach for Integrating Biological Heterogeneous Data Based on Connecting Ontology. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 600–607, https://doi.org/10.1109/BIBM52615.2021.9669700 (2021).
37. LePendu, P., Musen, M. A. & Shah, N. H. Enabling enrichment analysis with the human disease ontology. *Journal of Biomedical Informatics* **44**, S31–S38, https://doi.org/10.1016/j.jbi.2011.04.007 (2011).
38. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330–D338, https://doi.org/10.1093/nar/gky1055 (2019).
39. Cooper, L. *et al.* The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology* **54**, e1, https://doi.org/10.1093/pcp/pcs163 (2013).
40. Pan, Q. *et al.* Trait ontology analysis based on association mapping studies bridges the gap between crop genomics and phenomics. *BMC Genomics***20**, https://doi.org/10.1186/s12864-019-5812-0 (2019).
41. Evangelista, J. E. *et al.* Toxicology knowledge graph for structural birth defects. *Communications Medicine* **3**, https://doi.org/10.1038/s43856-023-00329-2 (2023).
42. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research* **47**, D1018–D1027, https://doi.org/10.1093/nar/gky1105 (2019).
43. Evangelista, J. E. *et al.* SigCom LINCS: data and metadata search engine for a million gene expression signatures. *Nucleic Acids Research* **50**, W697–W709, https://doi.org/10.1093/nar/gkac328 (2022).
44. Halip, L. *et al.* Exploring drugcentral: from molecular structures to clinical effects. *Journal of Computer-Aided Molecular Design* **37**, 681–694, https://doi.org/10.1007/s10822-023-00529-x (2023).
45. Lachmann, A. *et al.* Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic Acids Research* **47**, W571–W577, https://doi.org/10.1093/nar/gkz393 (2019).
46. Messina, A. *et al.* BioGrakn: A knowledge graph-based semantic database for biomedical sciences. *Advances in Intelligent Systems and Computing* **611**, 299–309, https://doi.org/10.1007/978-3-319-61566-0_28 (2018).
47. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**, D480–D489, https://doi.org/10.1093/nar/gkaa1100 (2021).
48. Putman, T. E. *et al.* The monarch initiative in 2024:ānānalytic platform integrating phenotypes, g enesānd diseasesācross species. *Nucleic Acids Research* **52**, D938–D949, https://doi.org/10.1093/nar/gkad1082 (2024).
49. Bult, C. J. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Research* **47**, D842–D848, https://doi.org/10.1093/nar/gky1056 (2019).
50. Boudin, M., Diallo, G., Drancé, M. & Mougin, F. The OREGANO knowledge graph for computational drug repurposing. *Scientific Data***10**, https://doi.org/10.1038/s41597-023-02757-0 (2023).
51. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data***10**, https://doi.org/10.1038/s41597-023-01960-3 (2023).
52. Sima, A. C. *et al.* Enabling semantic queries across federated bioinformatics databases. *Database* **2019**, baz106, https://doi.org/10.1093/database/baz106 (2019).
53. Kamran, A. B. & Naveed, H. GOntoSim: a semantic similarity measure based on LCA and common descendants. *Scientific Reports***12**, https://doi.org/10.1038/s41598-022-07624-3 (2022).
54. Ochs, C. *et al.* An empirical analysis of ontology reuse in BioPortal. *Journal of Biomedical Informatics* **71**, 165–177, https://doi.org/10.1016/j.jbi.2017.05.021 (2017).
55. Carlson, C. *et al.* The World Health Organization's Disease Outbreak News: A retrospective database. *PLOS Glob Public Health* **3**, e0001083, https://doi.org/10.1371/journal.pgph.0001083 (2023).
56. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112, https://doi.org/10.5555/2969033.2969173 (2014).
57. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**, https://doi.org/10.1002/widm.1249 (2018).
58. Azamfirei, R., Kudchadkar, S. R. & Fackler, J. Large language models and the perils of their hallucinations. *Critical Care* **27**, https://doi.org/10.1186/s13054-023-04393-x (2023).
59. Miller, G. A. WordNet: A lexical database for English. *Communications of the ACM* **38**, 39–41, https://doi.org/10.1145/219717.219748 (1995).
60. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, 2019).
61. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL-HLT 2019 Conference Proceedings*, vol. 1, 4171–4186, https://doi.org/10.18653/v1/N19-1423 (2019).
62. Korst, J., Pronk, V., Barbieri, M. & Consoli, S. Introduction to classification algorithms and their performance analysis using medical examples. In Consoli, S., Recupero, D. R. & Petkovic, M. (eds.) *Data Science for Healthcare: Methodologies and Applications*, 39–73, https://doi.org/10.1007/978-3-030-05249-2_2 (Springer Nature, 2019).
63. Lee, J. *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240, https://doi.org/10.1093/bioinformatics/btz682 (2020).
64. Deka, P., Jurek-Loughrey, A. & P, D. Evidence extraction to validate medical claims in fake news detection. In Traina, A. *et al.* (eds.) *Health Information Science*, 3–15, https://doi.org/10.1007/978-3-031-20627-6_1 (Springer Nature Switzerland, Cham, 2022).
65. Gong, X., Jiang, J., Duan, Z. & Lu, H. A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC Bioinformatics* **19**, https://doi.org/10.1186/s12859-018-2064-y (2018).
66. Zhou, Y. *et al.* A Short-Text Similarity Model Combining Semantic and Syntactic Information. *Electronics* **12**, https://doi.org/10.3390/electronics12143126 (2023).
67. Callahan, T. J. *et al.* An open source knowledge graph ecosystem for the life sciences. *Scientific Data* **11**, https://doi.org/10.1038/s41597-024-03171-w (2024).
68. Cavalleri, E. *et al.* An ontology-based knowledge graph for representing interactions involving RNA molecules. *Scientific Data* **11**, https://doi.org/10.1038/s41597-024-03673-7 (2024).
69. Gangemi, A. Ontology design patterns for semantic web content. In *Proceedings of the 4th International Conference on The Semantic Web*, ISWC'05, 262-276, https://doi.org/10.1007/11574620_21 (Springer-Verlag, Berlin, Heidelberg, 2005).
70. Shimizu, C., Hirt, Q. & Hitzler, P. A protégé plug-in for annotating OWL ontologies with OPLa. *Lecture Notes in Computer Science* **11155 LNCS**, 23–27, https://doi.org/10.1007/978-3-319-98192-5_5 (2018).
71. Cowell, L. G. & Smith, B.*Infectious Disease Ontology* (Springer Berlin Heidelberg, https://doi.org/10.1007/978-1-4419-1327-2_19 2010).
72. Neil Otte, J., Beverley, J. & Ruttenberg, A. BFO: Basic Formal Ontology. *Applied Ontology* **17**, 17–43, https://doi.org/10.3233/AO-220262 (2022).
73. Jackson, R. *et al.* OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database* **2021**, https://doi.org/10.1093/database/baab069 (2021).

74. Shen, W., Wang, J. & Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* **27**, 443–460, https://doi.org/10.1109/TKDE.2014.2327028 (2015).
75. Krauer, F. & Schmid, B. V. Mapping the plague through natural language processing. *Epidemics* **41**, https://doi.org/10.1016/j.epidem.2022.100656 (2022).
76. Soille, P. *et al*. A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Generation Computer Systems* **81**, 30–40, https://doi.org/10.1016/j.future.2017.11.007 (2018).
77. European Commission, Joint Research Centre (JRC). Epidemic Information Extraction from WHO Disease Outbreak News, European Data Portal, Joint Research Centre Data Catalogue, https://doi.org/10.2905/89056048-7f5d-4d7c-96ad-f99d1c0f6601 (2024).
78. Kirstein, F. *et al*. Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11685 LNCS**, 192–204, https://doi.org/10.1007/978-3-030-27325-5_15 (2019).
79. Kirillovich, A. & Nikolaev, K. Adapting the LodView RDF Browser for Navigation over the Multilingual Linguistic Linked Open Data Cloud. In *2022 IEEE 9th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2022*, 143–149, https://doi.org/10.1109/SETIT54465.2022.9875628 (2022).
80. Griset, P. & Schafer, V. Hosting the world wide web consortium for Europe: From CERN to INRIA. *History and Technology* **27**, 353–370, https://doi.org/10.1080/07341512.2011.604177 (2011).
81. Camarda, D. V., Mazzini, S. & Antonuccio, A. LodLive, exploring the Web of Data. In *ACM International Conference Proceeding Series*, 197–200, https://doi.org/10.1145/2362499.2362532 (2012).
82. Gangemi, A. & Presutti, V. Ontology design patterns. In *Handbook on Ontologies*, 221–243, https://doi.org/10.1007/978-3-540-92673-3_10 (International Handbooks on Information Systems, 2009).
83. Weibel, S. L. & Koch, T. The Dublin core metadata initiative: Mission, current activities, and future directions. *D-Lib Magazine***6**, https://doi.org/10.1045/december2000-weibel (2000).
84. Abbood, A., Ullrich, A., Busche, R. & Ghozzi, S. EventEpi-A natural language processing framework for event-based surveillance. *PLoS Computational Biology* **16**, https://doi.org/10.1371/journal.pcbi.1008277 (2020).
85. Consoli, S., Reforgiato Recupero, D. & Petkovic, M. (eds.) *Data Science for Healthcare-Methodologies and Applications*https://doi.org/10.1007/978-3-030-05249-2 (Springer Nature, 2019).
86. Biderman, S. *et al*. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, vol. 202, 2397–2430, https://doi.org/10.5555/3618408.3618510 (2023).
87. Dong, Q. *et al*. A survey on in-context learning. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1107–1128, https://doi.org/10.18653/v1/2024.emnlp-main.64 (Association for Computational Linguistics, Miami, Florida, USA, 2024).

Acknowledgements

We would like to thank the colleagues of the Digital Health Unit (JRC.F7) at the Joint Research Centre of the European Commission for helpful guidance and support. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission. This research did not receive any specific grant from funding agencies in the public, commercial, or nonprofit sectors. We would like to acknowledge the GPT@JRC initiative for providing access to the LLMs used in this study. We extend our gratitude also to the JRC Big Data Analytics Platform for providing secure access to a variety of open-source AI models. Finally, we would like to thank the colleagues contributing to the WHO Epidemic Intelligence from Open Sources (EIOS) initiative for the helpful suggestions and support during the development of this work.

Author contributions

S.C. conceptualized the work, developed the pipeline, and created the knowledge graph release. P.V.M., N.I.S. and M.C. conceptualized and supervised the work. L.B. and P.C. edited the original manuscript and analysed the results. N.S. helped in developing the software. I.B., L.S. and L.O. conceived the data collection and curated the dataset. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.C.