An LLM-Based Pipeline for Constructing Knowledge Graphs from Activity Records of Infectious Disease Patients

Shintaro Yamada[1,*], Yasunori Yamamoto[2], Takashi Okumura[3] and Ikki Ohmukai[1]

[1]The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
[2]National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, JAPAN
[3]Kitami Institute of Technology, 165 Koen-cho, Kitami, Hokkaido 090-8507, JAPAN

Abstract

Large language models (LLMs) have increasingly been used to construct knowledge graphs from natural-language text, but controlling the structural and terminological consistency of generated data remains challenging. This study proposes a schema-guided pipeline in which a Shape Expressions (ShEx) schema is semi-automatically derived from an existing ontology and RDF data, supplied to an LLM to guide RDF generation, and subsequently reused without modification for validation. We demonstrate the pipeline using the COVID-19 Infection Risk Ontology (CIRO) and synthetic activity histories designed to simulate data collected through COVID-19 epidemiological interviews. The dataset comprises approximately 30 days of activity histories for 78 fictional patients and 17,876 extracted events. Validation used a source-preserving profile and a stricter profile requiring location and complete temporal information for CIRO-based risk inference. All evaluated shapes achieved full conformance under the source-preserving profile, while strict-profile validation revealed substantial gaps associated mainly with missing source information and representation or ontology limitations. Among the nonconforming events, 801 (5.0%) showed at least one indication of an LLM-attributable error. These results suggest that ShEx can provide explicit, reusable structural constraints for LLM-based knowledge graph construction while also supporting post-generation validation and helping distinguish generation errors from limitations in source information and ontology coverage.

Keywords

schema-guided generation, infection risk ontology, knowledge graph construction, large language models, Shape Expressions

CEUR-WS.org/Vol-4316/pd-short-9.pdf

1. Introduction

The automatic construction of knowledge graphs from natural-language text requires highly accurate natural language processing for tasks such as knowledge extraction and relation inference. Large language models (LLMs) have recently been applied to these tasks. However, controlling the structural and terminological consistency of LLM-generated data remains challenging, and methods for achieving such control with high accuracy are still under development. In this study, we aim to construct knowledge graphs suitable for infection-risk inference by semi-automatically generating Shape Expressions (ShEx) from an existing ontology and providing both the ontology and the generated ShEx schemas to an LLM and subsequently reusing the same schemas, unmodified, to validate the result.

During the coronavirus disease 2019 (COVID-19) pandemic, public health nurses conducted manual interviews with infected individuals to identify close contacts and monitor infection trends. However, the rapid increase in the number of infections caused the workload associated with conventional epidemiological investigations to exceed operational capacity, placing severe pressure on the human resources of public health centers. This study proposes a pipeline for the semi-automatic construction,

IJCKG 2026 Posters, Demos and Workshops, November 19–21, 2026, Bangkok, Thailand
*Corresponding author.
sturla@l.u-tokyo.ac.jp (S. Yamada); yy@dbcls.rois.ac.jp (Y. Yamamoto); taka@wide.ad.jp (T. Okumura); i2k@l.u-tokyo.ac.jp (I. Ohmukai)
0000-0003-4412-7508 (S. Yamada); 0000-0002-6943-6887 (Y. Yamamoto); 0000-0001-8643-2864 (T. Okumura); 0000-0002-3276-3753 (I. Ohmukai)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

using an LLM, of knowledge graphs that are intended to support the identification of close contacts and the inference of infection risks from natural-language descriptions of the activities of patients with COVID-19. We demonstrate the pipeline on 17,876 events extracted from synthetic activity histories and evaluate the effectiveness of the proposed method.

2. Related Work

In Japan, previous studies have proposed methods for structuring administrative press releases concerning the movements of individuals with COVID-19 in Resource Description Framework (RDF) format [1] and developed an ontology for COVID-19 with a focus on places and spatial conditions [2]. In particular, the latter ontology enables the inference and retrieval of infection risks associated with the "Three Cs," a term used in Japan to describe three types of situations that increase the risk of COVID-19 transmission: closed spaces, close-contact settings, and crowded places.

Outside Japan, previous work has extracted knowledge about COVID-19 from publications and represented it as a network based on causal relationships [3]. Other studies have proposed COVID-19 ontologies from a medical perspective [4], as well as ontologies developed from a public health perspective [5]. Among the ontologies proposed in Japan and elsewhere, this study particularly draws on the COVID-19 Infection Risk Ontology (CIRO), which was developed in Japan [2]. This choice reflects the fact that the patient activity-history data examined in this study were designed to represent infections occurring in Japan.

The use of knowledge graphs and other graph-based structures has also attracted increasing attention in public health research on COVID-19. Examples include an attempt to conduct COVID-19 simulations using a graph-based structure [6] and a study that applied natural language processing (NLP) to construct and visualize a knowledge graph from case reports [7]. These studies demonstrate the utility of knowledge graphs for risk inference from a public health perspective.

Shape Expressions (ShEx) is a language for specifying structural constraints on RDF data [1]. It describes constraints on nodes and triples and is commonly used to validate RDF data structures. ShEx has also been used not only for validation but also for sharing data models and providing guidelines for data editing. Thornton et al. [8] reported that, in the context of Wikidata quality management, ShEx schemas facilitate both consensus building regarding data descriptions and the establishment of curation guidelines. Building on this view of ShEx as a set of guidelines, this study applies ShEx to the RDF data generation process.

In relation to these previous studies, the proposed method differs in two principal respects. First, the constraints imposed during RDF generation are expressed in an explicitly verifiable form: the ontology defines semantic relationships, while ShEx specifies structural requirements. Second, the ShEx schema provided as guidance during data generation is subsequently reused without modification as a validation schema. The proposed method is designed to support reliable knowledge extraction by an LLM from activity-history data through the generation of ShEx schemas based on an ontology. Consequently, even when other ontologies proposed in previous studies are used, the overall pipeline remains unchanged. We therefore expect that the proposed pipeline can support knowledge extraction from natural-language text using ontologies developed for a wide range of purposes, extending beyond the domain of infectious diseases.

3. Schema-Guided Knowledge Graph Construction Pipeline

Figure 1 provides an overview of the pipeline proposed in this study. Because the use of raw data obtained through interviews with actual patients was difficult owing to privacy and ethical considerations, we first used a large language model (LLM) to generate fictional patients' activity histories. To simulate data collected through actual interviews, all activity histories were generated as natural-language narratives.

[1] https://shex.io/shex-semantics/

The LLM was then used to segment these narratives into activities by time interval and convert them into structured JSON data.

Next, the CIRO ontology and synthetic data constructed on the basis of the ontology were provided to the LLM to generate a ShEx schema. Finally, the generated ShEx schema, the CIRO ontology, and the fictional patients' activity histories structured in JSON format were provided to the LLM to generate RDF data. Each stage of the pipeline is described in detail in the following sections.

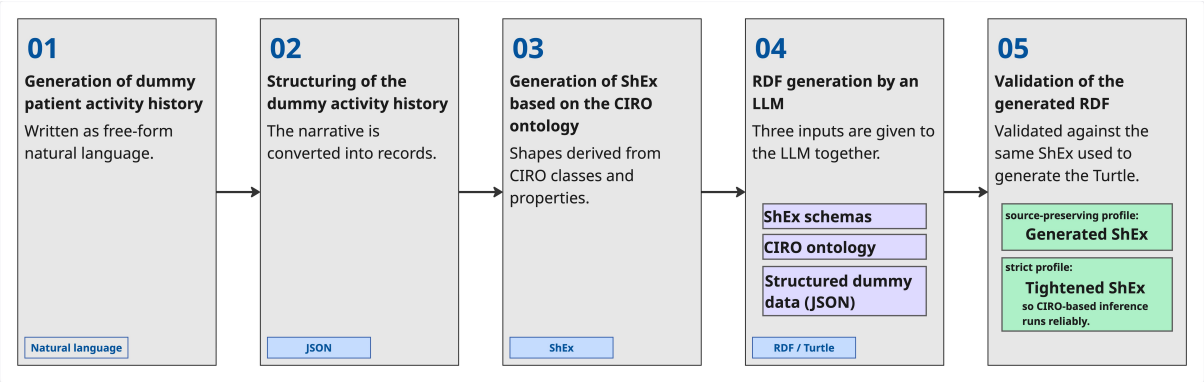


Figure 1: An Overview of the Proposed Pipeline

3.1. Patients' Activity History Data

Interviews concerning COVID-19 require the detailed collection of patients' activity histories over a defined period spanning the pre-symptomatic, symptom-onset, and post-onset phases. Given the characteristics of the disease, these histories provide essential information for identifying sources and routes of infection, as well as individuals who may have had close contact with an infected patient. However, activity histories are generally based on information obtained directly by public health nurses through interviews and are therefore typically recorded as unstructured natural-language text.

Patient activity-history data are also required in this study. Such data constitute sensitive personal information and must be handled with particular care because of privacy and ethical considerations. Accordingly, to evaluate the proposed method in a setting that reflects the operational environment of a public health center, we used an LLM to generate synthetic patient activity-history data. The generated data include basic patient information, such as name, age, sex, place of residence, occupation, and household composition, as well as activity histories covering a defined period before and after symptom onset. The activity histories contain information such as dates and times, places visited, accompanying persons, and mask use. By avoiding the use of actual patient information, this approach enables research on structuring interview data while addressing privacy and ethical concerns. The generated synthetic data were also reviewed by multiple public health experts, who confirmed that they contained no major inconsistencies with respect to infection mechanisms or medical plausibility. The procedure used to generate the patient data is described below.

1. Instruct the LLM to generate a patient profile containing basic attributes, including name, age, sex, address, occupation, and place of residence. If a public health expert has already provided a fictional profile, omit this step and use that profile directly.
2. Instruct the LLM to generate an activity history based on the generated or provided profile, including dates and times, visited locations, accompanying individuals, and mask use.
3. Consult public health experts to assess whether the generated data appropriately reflect transmission mechanisms and the characteristics of the variants circulating during the target period. Revise the relevant portions as necessary, also using the LLM to assist with these revisions.

Using the prompt shown in Figure 2, we generated the patient data presented in Table 1 with ChatGPT-4o. Because an interactive agent was used, generating an entire month of data in a single response

Generate a highly detailed activity history for the person described in the provided profile. Follow these rules:

1. Generate a complete activity history for each day from April 2 to April 30, 2022, using intervals of 15 minutes to one hour.
2. Ensure that the activity history is consistent with the provided profile.
3. Include activities that involve a risk of COVID-19 infection.
4. Describe each activity interval in detail using at least 50 Japanese characters.
5. Explicitly state the time of each activity.
6. When describing the use of public transportation, such as trains or buses, use actual nearby station names and other relevant locations consistent with the provided profile.
7. Do not use code-generation functionality.
8. Output the activity history in TSV format, with the time interval at the beginning of each row.

Figure 2: Prompt used to generate synthetic patient activity histories.

was impractical. We therefore adjusted the amount of data generated at one time, typically requesting activity histories for one to three days per response. Through this iterative process, we ultimately created approximately 30 days of patient data, covering April 2–30, 2022, for 78 fictional patients.

Table 1
Activity history of a patient on April 23, 2022 (excerpt)

Time	Activity
13:00	Measured his body temperature again (36.4°C); no problems were observed. He spent time watching television at a low volume, checked the weather forecast, and organized his schedule for the following week.
15:00	Took a short nap. He woke after approximately 30 minutes and felt almost no physical heaviness. After drinking some water, he performed light stepping exercises.
16:00	Checked the COVID-19 situation in nearby areas on the news. Conscious of the risk of reinfection, he reviewed how frequently he washed his hands and ventilated the room.
17:30	Ate udon noodles, simmered vegetables, and chilled tofu for dinner. He could taste the seasoning and finished the entire meal. He also remembered to take his medication after dinner.
18:30	Watched a historical drama on television. He felt that his improved ability to concentrate allowed him to enjoy more activities. He did not cough.

As described above, this study assumes that actual patients would provide their activity histories in natural language. The generated data therefore include natural variation in the length and level of detail of the descriptions. We organized these data in Google Sheets, with one patient per row and each column containing that patient's activity history for a single day. The resulting tabular data were then converted into JSON format using an LLM, and the JSON data were subsequently used to construct the knowledge graph[2].

3.2. Semi-Automatic Generation of Shape Expressions

Next, we aimed to ensure the consistency of the RDF data constructed from the patient data by using Shape Expressions (ShEx). For this purpose, a ShEx schema was generated semi-automatically using an LLM. The procedure was as follows.

1. Provide the COVID-19 Infection Risk Ontology (CIRO) and RDF data created using the ontology (PLOD) to an LLM and generate a ShEx schema.
2. Validate the existing PLOD data against the generated ShEx schema.
3. If the generated schema does not adequately conform to the structure of the existing PLOD data, revise the ShEx schema while leaving the PLOD data unchanged. Repeat Steps 2 and 3 until the schema appropriately captures the structure of the PLOD data.

[2]Prompt file can be accessed from https://github.com/s-geirmundur/IJCKG_2026/blob/main/Excel2JSON_prompt_v2.yaml

We used ChatGPT-5.6 Sol for this process. The synthetic PLOD data and CIRO were provided simultaneously, and the model was instructed to generate a ShEx schema according to the YAML-formatted prompt.[3] The resulting ShEx schema and PLOD data were then provided to ChatGPT-5.5 for consistency validation. Whenever an error was identified during validation, the model was instructed to generate a revised ShEx schema in which the relevant issue had been corrected. This process was repeated until no further errors occurred, and the resulting schema was adopted as the final ShEx schema (Figure 3)[4].

```
<EventShape> EXTRA rdf:type rdfs:label rdfs:comment {
  rdf:type [ schema:Event ] ;
  rdfs:comment . * ;
  plod:time @<AnyIntervalShape> ;
  schema:location @<PlaceShape> ? ;
  plod:agent @<AgentShape> + ;
  plod:action @<ActionValueShape> * ;
  plod:context @<BehavioralContextValueShape> * ;
  plod:situationOfAction . {0} ; plod:situationOfSpace . {0} ;
  plod:situationOfActivity . {0} ; plod:isSituationOf . {0}
}

<AnyIntervalShape> @<CompleteIntervalShape> OR @<PartialIntervalShape>

<CompleteIntervalShape> EXTRA rdf:type rdfs:label rdfs:comment {
  rdf:type [ time:Interval ] ;
  time:hasBeginning @<InstantShape> ;
  time:hasEnd @<InstantShape> ;
  time:hasDuration @<DurationShape>
}

<PartialIntervalShape> @<IntervalCoreShape> AND NOT @<CompleteIntervalShape>
```

Figure 3: Excerpt of the ShEx schema for EventShape, showing complete and partial temporal interval representations

In the ShEx schema, `ActionValueShape`, `BehavioralContextValueShape`, and `SpatialContextValueShape` are defined as sets of permitted IRIs rather than node shapes. Specifically, they enumerate eight, four, and two IRIs, respectively, and are checked as value constraints within `EventShape` and `SituationShape`. Because they define no independent population of target nodes, no shape-level conformance rate applies to them; their presence in the graph is instead assessed through the coverage measures introduced in Section 4.

We performed ShEx validation using two profiles. The first, the source-preserving profile, used the same ShEx schema that guided the LLM during RDF construction. This profile assesses whether the RDF follows the required structure while representing the information available in the source, without requiring the LLM to supply missing information. The synthetic activity histories represent diverse patients and vary in granularity and expression. For example, some activities have explicit start and end times, whereas others do not. In this setting, allowing the LLM to fill in unspecified details could undermine the reliability of subsequent infection risk assessment. Missing information alone is therefore not treated as nonconformance under this profile.

The second, the strict profile, assesses whether events contain the information required for CIRO-based risk inference. It requires a location, start time, end time, and duration. This profile differs from the source-preserving profile in only two respects. First, the event shape requires exactly one location. CIRO uses location information represented through the location shape to derive spatial risk, so events

[3]Prompt file can be accessed from https://github.com/s-geirmundur/IJCKG_2026/blob/main/RDF2ShEx_prompt_v2.yaml
[4]ShEx file can be accessed from https://github.com/s-geirmundur/IJCKG_2026/blob/main/PLOD_conversion_CIRO_2026_ver2-0.shex

without a location cannot support this inference. Second, the temporal shape requires both interval endpoints and a duration derived from them, because incomplete intervals do not adequately support the intended temporal risk inference. All other definitions are identical across the two profiles.

For the source-preserving profile, the same ShEx schema therefore plays two distinct roles in the proposed pipeline. During RDF generation, it is supplied to the LLM as an explicit structural specification. After generation, the same schema is reused without modification as a validation specification to test whether the resulting RDF conforms to those structural requirements.

3.3. LLM-Based Knowledge Graph Construction Using ShEx

Using the ShEx schema created in the preceding section and the patient activity-history data in JSON format, we constructed a knowledge graph based on the prompt.[5] Codex GPT-5.6 Sol was used for this process. As a result, the RDF data shown in Figure 4 were generated. The data represent each patient activity as a time-specific event, together with information such as the date, time, location, and type of activity.

For activities not defined in CIRO, such as reading a newspaper, the LLM did not arbitrarily interpret the activity or introduce new vocabulary. Instead, it reproduced the patient's original description as an rdfs:comment. The prompt did not explicitly instruct the model to adjust the granularity of event descriptions to conform to the specified data format. It only instructed the model not to modify or extend the definitions in the provided ontology at its own discretion. This behavior is consistent with the constraints specified in the ShEx schema supplied to the LLM together with the prompt. More generally, the pipeline preserves, in `rdfs:comment`, the source text from which each event was segmented together with the rationale for its extraction decisions. These annotations are what makes the cause classification in Section 4 possible.

```
data:P107 a schema:Person ;
    rdfs:label "Sae Horiuchi" .          # a synthetic patient

data:P107_event_20220402_10 a schema:Event ;
    plod:agent      data:P107 ;
    plod:time       data:P107_time_20220402_10 ;
    schema:location data:P107_place_20220402_10 ;
    plod:action     plod:eat ;            # DropletReachableAction
    plod:context    plod:relax ;          # BehavioralRiskContext -> on the Event
    rdfs:comment
        "evidence for plod:eat: coffee",
        "evidence for plod:relax: the source mentions resting / chatting",
        "ventilation / interpersonal distance recognised but not representable in CIRO",
        "source segment: 14:30-15:15, a break at a cafe near the park; The kids drank fruit juice, and the adults ordered coffee. The
      place was crowded, and being seated so close to other customers made it a bit uncomfortable." .

data:P107_place_20220402_10 a plod:CoffeeShop ;
    rdfs:label "cafe" .

data:P107_situation_20220402_10 a plod:Situation ;
    plod:isSituationOf data:P107_place_20220402_10 ;
    plod:time          data:P107_time_20220402_10 ;
    plod:context       plod:crowded ;     # SpatialRiskContext -> on the Situation
    rdfs:comment "evidence for plod:crowded: the source mentions crowding" .

data:P107_time_20220402_10 a time:Interval ;   # conforms to CompleteIntervalShape
    time:hasBeginning data:P107_instant_20220402_10_begin ;
    time:hasEnd       data:P107_instant_20220402_10_end ;
    time:hasDuration  data:P107_duration_20220402_10 .

data:P107_instant_20220402_10_begin a time:Instant ;
    time:inXSDDateTime "2022-04-02T14:30:00"^^xsd:dateTime .
data:P107_instant_20220402_10_end a time:Instant ;
    time:inXSDDateTime "2022-04-02T15:15:00"^^xsd:dateTime .
data:P107_duration_20220402_10 a time:TemporalDuration ;
    time:numericDuration 45 ; time:unitType time:unitMinute .
```

Figure 4: Generated RDF for a single activity (excerpt; prefix declarations omitted). The behavioral context `plod:relax` is attached to the event, whereas the spatial context `plod:crowded` is attached to the situation, following CIRO. Each assignment records its source-text evidence in `rdfs:comment`, as does the decision not to represent ventilation and interpersonal distance, for which CIRO provides no corresponding individuals.

[5]Prompt file can be accessed from https://github.com/s-geirmundur/IJCKG_2026/blob/main/JSON2RDF_prompt.yml

4. Validation of the Generated Knowledge Graph

We evaluated the generated RDF against CIRO and the ShEx schemas using shape-level conformance and event-level coverage. Here, ShEx conformance evaluates structural conformance to the specified RDF constraints; it should not be interpreted as a direct measure of semantic correctness. An RDF graph may satisfy the ShEx schema while still containing an incorrect entity, event, or vocabulary mapping. Conformance was calculated over generated nodes targeted by each shape; structural coverage used all 17,876 events as the denominator. Conditional coverage was restricted to events with relevant source evidence or a prerequisite element. Event nonconformance was classified using retained `rdfs:comment` annotations and, for location mapping, a vocabulary derived from generated Place labels. The rules cover only the four strict-profile requirements and exclude over-assignment of actions and risk contexts; the identified errors therefore represent a lower bound.

Table 2
ShEx conformance rates

	Source-preserving		Strict	
Shape	n	%	n	%
Event	17,876	100.0	1,994	11.1
Situation	4,265	100.0	1,878	44.0
Interval	17,876	100.0	7,660	42.8

All seven evaluated shapes achieved full conformance under the source-preserving profile, including CIRO's separation of behavioral contexts on events from spatial contexts on situations. Under the strict profile, Event, Situation, and Interval conformance decreased (Table 2); Instant, Duration, Place, and Agent remained fully conformant.

Table 3
Structural and conditional coverage of event elements

	Structural		Conditional	
Element	n	%	Denominator	%
`schema:location`	4,437	24.8	events with a location expression	86.3
`time:hasBeginning`	16,552	92.5	events with a time expression	99.8
`time:hasEnd`	7,696	43.0	events with an explicit time range	96.5
`time:hasDuration`	7,660	42.8	events with both endpoints	99.7
Corresponding `plod:Situation`	4,265	23.8	events with a location identified	96.1

Structural coverage uses all 17,876 events as the denominator. Conditional denominators are either events whose source text provides the relevant evidence or events that already contain a prerequisite element. `plod:action` and plod:context were retained by 6,674 (37.3%) and 1,547 (8.6%) events respectively; no conditional denominator is defined for them.

Table 3 shows high start-time coverage but limited location and complete-interval coverage, explaining the low strict-profile event conformance. The substantially higher conditional coverage indicates that source availability and prerequisite elements account for much of this gap.

Of the 15,882 nonconforming events, 801 (5.0%) showed at least one indication of an LLM-attributable error: 699 location mapping errors, 15 start-time mapping errors, and 129 segmentation errors, with overlapping categories deduplicated. The remaining 15,081 were attributed to insufficient source information, unsupported profile requirements, normalization limitations, or insufficient ontology coverage. The latter included 403 cases involving ventilation or interpersonal distance that the model recognized but CIRO could not represent. Omissions of explicitly stated durations can fall outside these rules, which classify missing durations solely by interval endpoints.

Mapping errors suggest inconsistent application of recognized mappings: `plod:Home` was assigned to only 1,095 of 1,609 events (68.1%) containing "return home." Segmentation errors introduced spurious

events when embedded clock times were mistaken for boundaries during JSON structuring; for example, a reference to a "meeting starting at 14:00" within a 12:15–14:00 activity incorrectly started a second event.

5. Conclusion

We proposed an LLM-based pipeline that uses CIRO and a semi-automatically derived ShEx schema to guide knowledge graph construction from patient activity histories, then reuses the schema unmodified for validation. Applied to 17,876 events from 78 synthetic patients, it achieved full conformance under the source-preserving profile, while strict-profile validation exposed location and temporal gaps. Under our classification rules, most nonconformance was attributed to source or representation limitations, with identifiable LLM-attributable errors in 5.0% of nonconforming events. These counts are a lower bound, and schema conformance guarantees neither consistent semantic mapping nor the soundness of supporting evidence.

A further limitation is that the present evaluation uses synthetic activity histories rather than records collected in actual epidemiological investigations. Real-world records may contain incomplete or ambiguous descriptions, missing or inconsistent timestamps, conflicting information across sources, and clinical or behavioral information that is not covered by the current ontology. Such cases may reduce both extraction accuracy and the usefulness of structurally conformant RDF for downstream risk inference. Evaluation on appropriately de-identified real-world data is therefore required to assess the generalizability of the pipeline. Future work will compare ShEx with SHACL, evaluate API-based batch processing, and extend CIRO to additional risk-relevant concepts.

Acknowledgments

This work was supported by the Japan Science and Technology Agency (JST), Research Institute of Science and Technology for Society (RISTEX), Solution-Driven Co-creative R&D Program for SDGs, under grant JPMJRS23I5.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 5, 5.5 (Anthropic) and GPT-6 Astra (OpenAI) for grammar correction, translation and assistance with drafting. The authors reviewed and revised the resulting content and took full responsibility for the final content of the paper.

References

[1] I. Ohmukai, Y. Yamamoto, M. Ito, T. Okumura, Tracing patient PLOD by mobile phones Mitigation of epidemic risks based on patient locational open data, in: 2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2020, pp. 283–286. URL: https://ieeexplore.ieee.org/document/9338506. doi:10.1109/WETICE49692.2020.00062.

[2] S. Egami, Y. Yamamoto, I. Ohmukai, T. Okumura, CIRO: COVID-19 infection risk ontology, PLOS ONE 18 (2023) e0282291. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0282291. doi:10.1371/journal.pone.0282291.

[3] D. Domingo-Fernández, S. Baksi, B. Schultz, Y. Gadiya, R. Karki, T. Raschka, C. Ebeling, M. Hofmann-Apitius, A. T. Kodamullil, COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology, Bioinformatics (Oxford, England) 37 (2021) 1332–1334. doi:10.1093/bioinformatics/btaa834.

[4] Y. He, H. Yu, E. Ong, Y. Wang, Y. Liu, A. Huffman, H.-h. Huang, J. Beverley, J. Hur, X. Yang, L. Chen, G. S. Omenn, B. Athey, B. Smith, CIDO, a community-based ontology for coronavirus

disease knowledge and data integration, sharing, and analysis, Scientific Data 7 (2020) 181. URL: https://www.nature.com/articles/s41597-020-0523-6. doi:10.1038/s41597-020-0523-6.

[5] B. Dutta, M. DeBellis, CODO: An Ontology for Collection and Analysis of Covid-19 Data, in: Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2020, pp. 76–85. URL: http://arxiv.org/abs/2009.01210. doi:10.5220/0010112500760085, arXiv:2009.01210 [cs].

[6] Y.-H. Wu, T. E. M. Nordling, CovSyn: an agent-based model for synthesizing COVID-19 course of disease and contact tracing data, 2025. URL: https://www.medrxiv.org/content/10.1101/2025.05.17.25327820v1. doi:10.1101/2025.05.17.25327820.

[7] J. Wang, K. Wang, J. Li, J. Jiang, Y. Wang, J. Mei, S. Li, Accelerating Epidemiological Investigation Analysis by Using NLP and Knowledge Reasoning: A Case Study on COVID-19, AMIA Annual Symposium Proceedings 2020 (2021) 1258–1267. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8075493/.

[8] K. Thornton, H. Solbrig, G. S. Stupp, J. E. Labra Gayo, D. Mietchen, E. Prud'hommeaux, A. Waagmeester, Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation, in: P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, K. Hammar (Eds.), The Semantic Web, Springer International Publishing, Cham, 2019, pp. 606–620. doi:10.1007/978-3-030-21348-0_39.