



Project Data-Highway
A Linked Open Data Platform for
E-Tourism in Sicily
CUP G69J18001180007

An LLM Toolchain for Open Data and NGSI-LD

Insights from the **Data-Highway** Project



UNIONE EUROPEA



REPUBBLICA ITALIANA



REGIONE SICILIANA



The Data-Highway Project

Goal:

- An **intelligent, interoperable** data infrastructure for smart tourism and cultural heritage
- By leveraging **NGSI-LD** (an ETSI standard for Linked Data), enable seamless **access** and **reuse** of **Linked Open Data**

Reality:

- **Unstructured Data:** Critical information is locked in **natural language** (*web pages, PDFs, reviews*)
- **Poorly Structured Data:** Existing **datasets** (*CSV, JSON*) are **disorganized**, inconsistent, and lack a common standard

Our target schema: NGSI-LD

Why NGSI-LD? (www.ngsild.org)

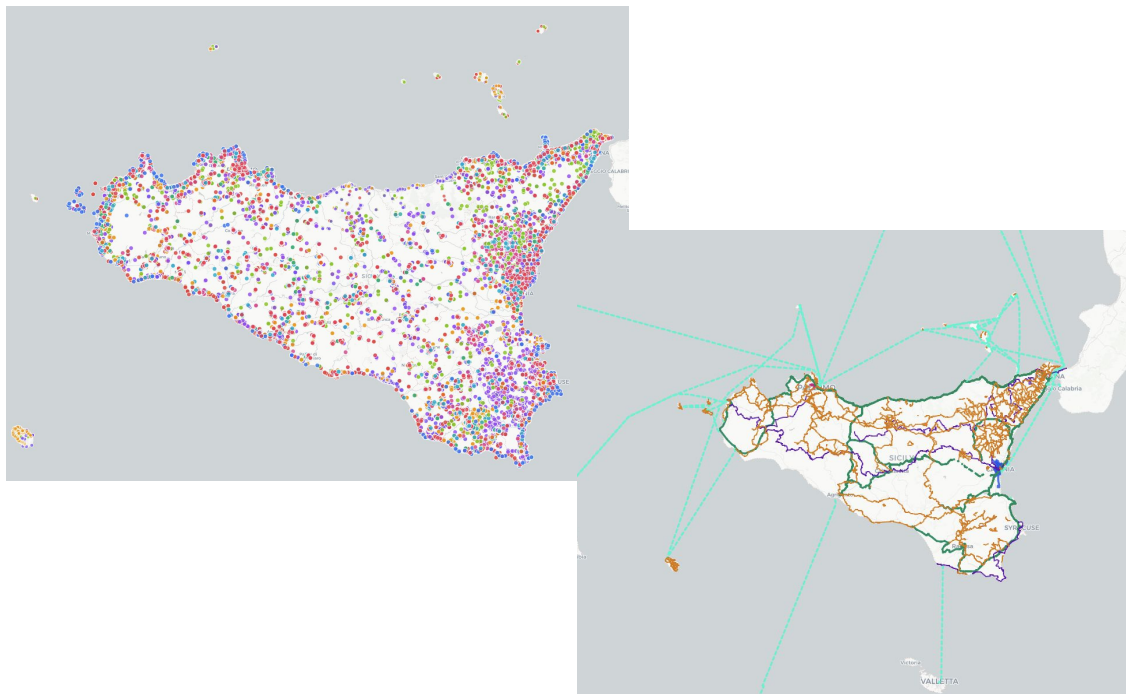
1. **Interoperability:** Creates a common language for diverse data sources
2. **Linked Data Native:** Built on top of JSON-LD, at the heart of Linked Data
3. **Enables Digital Twins:** Perfect for modeling real-world entities and their relationships

Giardino Bellini, Catania, Sicily



```
{
  "id": "urn:ngsi-ld:Garden:GiardinoBellini",
  "type": "Garden",
  "isLocatedAt": {
    "type": "Relationship",
    "object": "urn:ngsi-ld:City:Catania"
  },
  "location": {
    "type": "GeoProperty",
    "value": { ... }
  },
  ...
}
```

Ongoing Geographical Data Coverage



Types

+100 unique Entity types

Entities

+54,000 entities

Total length of public transport lines

+18,000 Km of lines covered

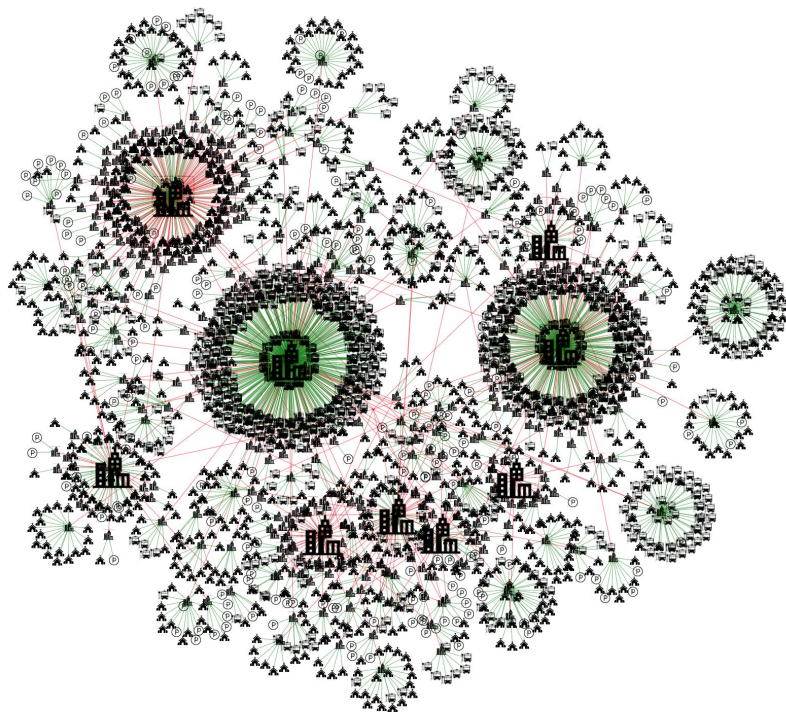
NGSI-LD represents a Property Graph

Relationships enable (distributed) graph and graph queries

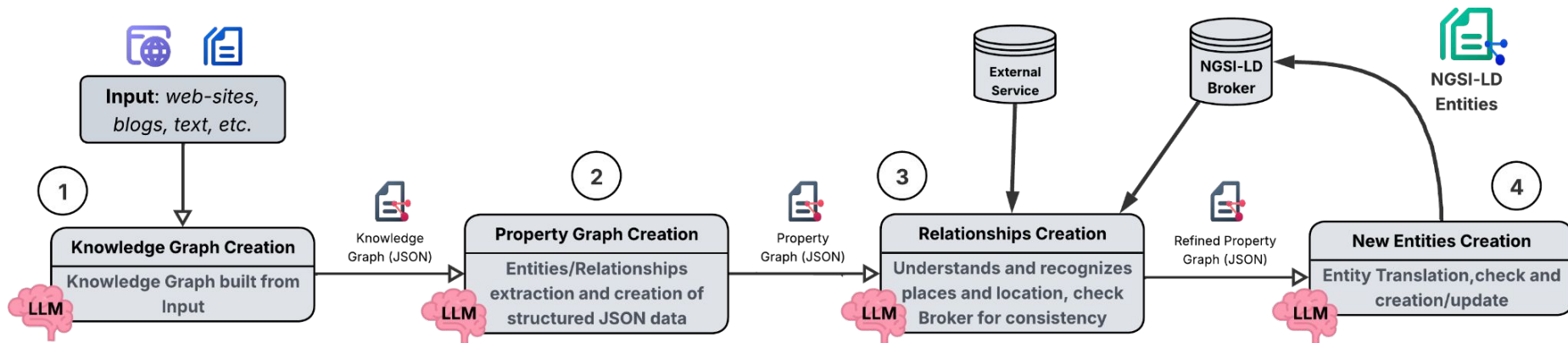
+54,000 Relationships between Entities

The graph shows the nine Sicilian Provinces as main nodes from which municipalities, points of interest, and events are articulated

(Less than 5% of the total available data shown)



Our Approach: The LLM Toolchain Architecture



Step 1: Knowledge Graph Creation

- Identifies facts and relationships among them
- Forces JSON output

Step 2: Property Graph Creation

- Secondary (e.g. timestamps, locations) facts are embedded as Properties

Step 3: Relationship Creation

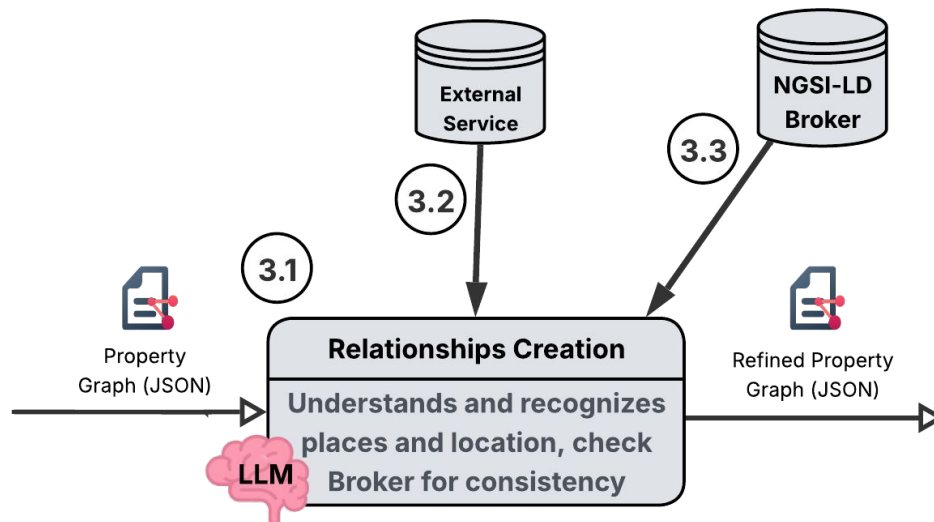
- A series of modules determines (e.g. via external ID matching) existing **target entities**

Step 4: New Entities Creation and NGSI-LD Translation

- New Entities are created **if empty target** entities exist
- **Translation** of JSON objects to **NGSI-LD Entities** done

Careful prompt engineering at each step

Example: Place ID Relationship Resolver



```
{BROKER_URL}/ngsi-ld/v1/entities?geoproperty=location&
geometry=Point&coordinates=[lng,lat]&georel=near;
maxDistance==<d>&type=<t1,t2,...>
```

Step 3.1: Normalize Location Name

- LLM understands the **location** and tries to **normalize** the **name**

Step 3.2: Check via Geoservice ID

- Module **queries Geoservice API** for a matching place name, returning an **external ID** if found
- The module queries the NGSI-LD broker for a matching target entity having the external ID

Step 3.3: Check via expanding GeoQuery

- If Step 3.3 doesn't match, an LLM worker returns **3 types** that can fit the actual Place, and an **NGSI-LD GeoQuery** is executed. An **increased distance** is used in each query



Step 2, Property Graph Creation: Example Prompt

****DateTime informations****

DO NOT CREATE entity of temporal information such as (date, day, month, year, time).
However, temporal information should be captured as properties of Event entities.

Ensure that it doesn't create Entities related to Date and Time

****Entity Identification**:**

- Extract all distinct entities mentioned in the Knowledge Graph
- Identify entity types: Event, Person, Artist, Band, <VenueType>, City, Organization, etc.
- Ensure entities are semantically distinct and not duplicated with variations
- Normalize entity names to canonical forms

Ensure that Entities are extracted from the Knowledge Graph, it should choose a certain type for that entity (from the Broker's type list).

****Main Entity Definition**:**

- Determine the main entity that should have direct or indirect connections to other entities
- Ensure this entity serves as the central hub of the knowledge graph

Ensure that the LLM defines a Main Entity that have Relationships to other Entities

****Relationship Identification**:**

- Determine explicit relationships between entities as stated in the Knowledge Graph
- Infer implicit relationships based on semantic context
- Create bidirectional relationships where appropriate (e.g., hasEvent, isLocatedAt, BUT NOT isLocateAt, isLocateIn)
- Include spatial relationships (isLocatedAt BUT NOT isLocateAt, isLocateIn)
- Include performative relationships (performsAt, features)

Ensure that the LLM has a better context on the main Relationship used and how to choose one, it also give some guidance on Relationship names to prevent hallucinations.

Example of Unstructured Text Inputs

INPUT 1

Christmas 2025 in Messina

This **December**, **Messina** invites you to celebrate the **Christmas season** with family and friends. The city will host several festive activities, including **games**, **movie projections**, and great **food**.

The celebrations begin on **December 24th** with the traditional "**Tombolata di Natale**." On Christmas Day, **December 25th**, families can enjoy **movie projections for children**. The holiday events will conclude on **December 30th** with an evening of **tastings** and **celebrations**.

INPUT 2

Catania Summer Fest Returns to Giardini Bellini

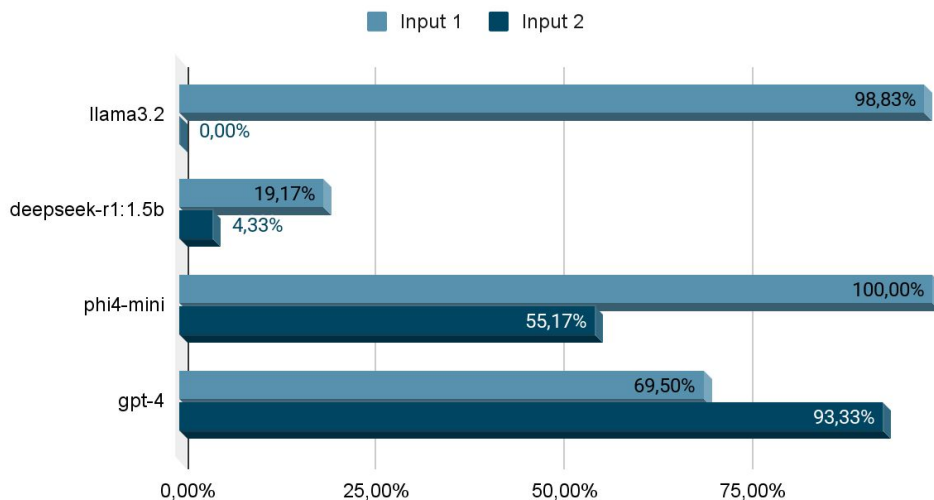
This season, the vibrant heart of Sicily beats at **Giardini Bellini** with the return of Catania Summer Fest. Running **from June 26th through October 11th**, the festival brings the city alive with sizzling sounds, local flavours, and unforgettable moments.

The festivities kick off on **June 26th at 18:15** with a welcome event featuring **DJs** spinning chill **music**. Later in the summer, on **July 12th at 19:00**, attendees can immerse themselves in local culture with a **Sicilian cooking demonstration** as part of the festival's cultural workshops.

Performance of different LLMs:

Step 3, Relationships Creation

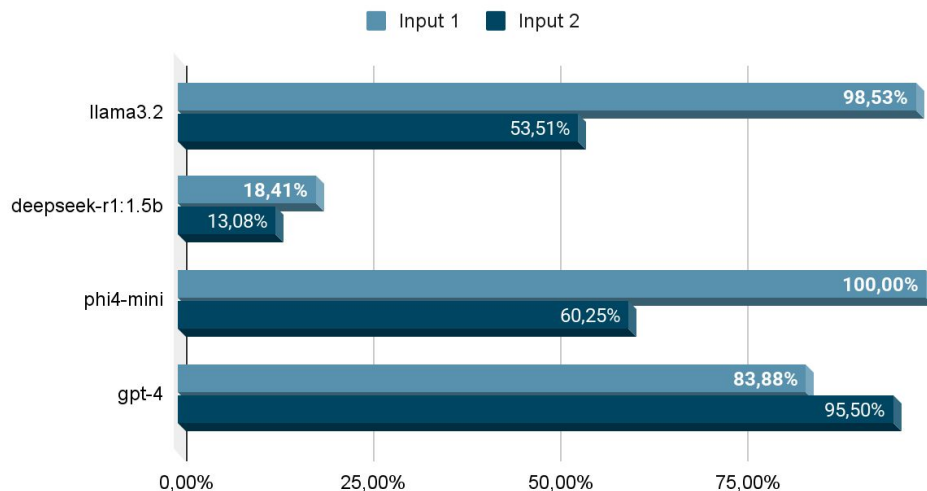
Percentage of correct Relationships by model



- **Deepseek hallucinated** during its **thought process**, evaluating the **name** of the place **incorrectly**. It also provided **types** to be used in queries that are **not related** to the actual **place**. This led to poor results
- **Llama3.2** heavily **hallucinated** on Input 2 during the **NGSI-LD translation**, replacing the type in the **id** (e.g. urn:ngsi-ld:Type:001) with the word **“Place”**. This led to poor results

Performance of different LLMs: Step 4, Entities Creation

Percentage of correct Entities by model



The correctness of the Entities Creation is based on a weighted linear combination of:

- Meeting the expected number of Entities
- Having the correct Property values for each entity
- Correct Relationships to target Entities

Final considerations

- Viable method involving **Knowledge Graph** to:
 - automate the **translation** of **natural-language data** into **structured JSON**
 - Translate into the interoperable **NGSI-LD format**
- Need for additional **disambiguation** services
 - crucial for data integrity and entity fusion
- While models like **gpt-4** showed **high reliability**, others like **Llama3.2** and **Deepseek** suffered from critical **hallucinations** (e.g., incorrect place names, non-existent URN) that led to poor results. Need to investigate with other/newer open-source models



Thank you for the attention!

Any Questions?

giuseppe.tropea@netsenseweb.com | www.netsenseweb.com

