Data-Highway: a Linked Open Data Platform for E-Tourism in Sicily

Giuseppe Tropea and Luca Longo
NetSense srl, Italy
https://www.netsenseweb.com
giuseppe.tropea@netsenseweb.com
luca.longo@netsenseweb.com

Andrea Detti
CNIT, Italy
Università di Roma Tor Vergata
andrea.detti@uniroma2.it

***Abstract*—Interoperable E-Tourism applications need two key ingredients: a large set of high-quality data about the territory and a developers-friendly API that is able to support complex queries on the data. The Data-Highway project leverages two consolidated technologies, Linked Open Data and NGSI-LD, together with a toolchain centred on LLMs, to create a tangible added value in Sicily in the tourism sector.**

I. Technologies for E-Tourism

For an ecosystem of E-Tourism applications that thrives and is self-sustainable, two key ingredients are needed: a large set of high-quality data about the region and a developers-friendly API that is able to support complex queries on the data. The companies and businesses operating in the tourism sector in Sicily, that are increasingly relying on information technology to offer new or better service, would benefit from a lower technological barrier when developing their Apps, thus increasing quality and competitiveness.

Still, offering curated data in a simple fashion to application developers is extremely challenging, and, generally speaking, a problem that has no solution in the more generic case. However, focusing on a specific sector (tourism) and on a specific region (Sicily) makes the problem more tractable, and it makes the Data-Highway project, funded by Regione Sicilia, able to create a tangible added value. The project is currently (June 2024) ongoing and it will end in 2025.

The project leverages two consolidated technologies: Linked Open Data (LOD) [1] and NGSI-LD [2]. NGSI-LD is a standard defined by ETSI for managing and exchanging context information. It incorporates principles from Linked Data, enabling the interconnection of different data sources and the creation of a web of data; the model revolves around entities (things in the real world or concepts) and their relationships. NGSI-LD also defines a standard API, which developers can use to create, update, retrieve, and delete *context information*: this is particularly important in smart cities, where diverse E-Tourism applications may need to work together.

We start from an architecture similar to DET [3], but our project's innovation is the usage of LLM (Large Language Models) to develop a toolchain for data cleansing and correlation that automatizes, to the extent that is possible, the updating and maintaining of a valuable data infrastructure.

II. Data Sources

Dati.gov.it is the national portal for open data in Italy. It serves as a comprehensive catalog of metadata related to open data released by various public administrations. The portal was initiated in 2011 and is managed by the Agency for Digital Italy (AgID) since 2015. By exploiting the LOD principles, the portal facilitates the search and reuse of data by making metadata more accessible via both a SPARQL and a CKAN interface.

However, the data released by the administrations is extremely heterogeneous and does not, in general, adhere to LOD principles. The goal of the project is to enrich the available data with context information coming from various sources (DBpedia, GeoNames, Wikidata, OpenStreetMap), thus mapping data into the NGSI-LD standardized format.

III. Data Gathering

Our toolchain periodically scans the 77k+ data catalogues ("packages") of dati.gov.it, and it applies some business logic to analyse the available location information of each package. Then, a mapping with GeoNames is performed, to extract the $\{$latitude, longitude$\}$ or the bounding box of the package. By doing so, a geo-fencing and pruning of the federated data sources is possible, thus limiting the coverage to Sicily only.

Dati.gov.it is regularly updated. This means that new packages can be added or removed. Data-Highway is designed to cope with this: if a new package is inserted, the steps above are performed incrementally. If a package is removed, the corresponding element is deleted from the collection that lives in our supporting mirror geo-database. Each package contains one or more urls that link to the actual data made available by the federated public administration, thus an online resource. Files can can have different formats. We are currently focusing on on csv, json and geojson files. Every file is retrieved and can now be passed on to an LLM-centric toolchain for analysis.

IV. LLM for Data Transformation

In this phase, the project is focused on the challenge of automating the process of analysis and creation of NGSI-LD entities, starting from structured csv, json and geojson documents. Through the support of an LLM, it is possible to

engineer one or more template prompts that facilitate the process. Currently, the LLM is tasked in extracting/inferring the entity type, in incrementally updating the NGSI-LD @context file, in detecting duplicates, and in reshaping the identifier and attributes of the entity itself, consistently mapping to NGSI-LD. We use the openAI's GPT models and APIs, currently experimenting with GPT-4-turbo and GPT-4o models and a temperature value of 0.7.

A. Extracting entity type

The first step is to extract an entity type. The LLM is given a list of entity types that are already present in the NGSI-LD broker, and the task is to reuse one of them in order to avoid redundancy with entity types that have the same meaning, or to instantiate a new type, if needed. The prompt is engineered around two main ideas: (1) find the entity type by analyzing the properties and values of the entity; (2) if you cannot extract the entity type from properties or the values use the filename to guess extract the best entity type.

B. Incrementally Updating the @context

Through this prompt's ruleset we use the LLM to maintain @context entries that have the same meaning but a different name. An example output fragment is provided below.

```
"@context": {
  "name": "https://schema.org/name",
  "address": "https://schema.org/address",
  "lng": "https://schema.org/longitude",
  "lon": "https://schema.org/longitude",
  "Nominativo": "https://schema.org/name",
  "Address": "https://schema.org/address",
  "Longitude": "https://schema.org/longitude"
}
```

C. Obtaining Relationships

Using the LLM we try to obtain relations between entities. For instance, feeding it a json as a text, that contains a list of dams in Sicily, like this:

```
{"diga_id":17,
"diga_nome":"POZZILLO",
"corso_acqua_nome":"SALSO (SIMETO)",
"utilizzo":"IRRIGUO;
            PRUDUZIONE ENERGIA ELETTRICA",
"ente_gestore":"E.N.E.L.",
"longitudine":14.5914590,
"latitudine":37.6604350,},

{"diga_id":18,
"diga_nome":"PRIZZI",
"corso_acqua_nome":"RAIA",
"utilizzo":"IRRIGUO;
            ACQUA POTABILE;
            PRUDUZIONE ENERGIA ELETTRICA",
"ente_gestore":"E.N.E.L.",
"longitudine":13.4046770,
"latitudine":37.7282200,},

{"diga_id":19,
"diga_nome":"RAGOLETO",
"corso_acqua_nome":"DIRILLO",
"utilizzo":"INDUSTRIALE;
            ACQUA POTABILE",
"ente_gestore":"RAFFINERIA DI GELA",
"longitudine":14.6963330,
"latitudine":37.1282220,}
```

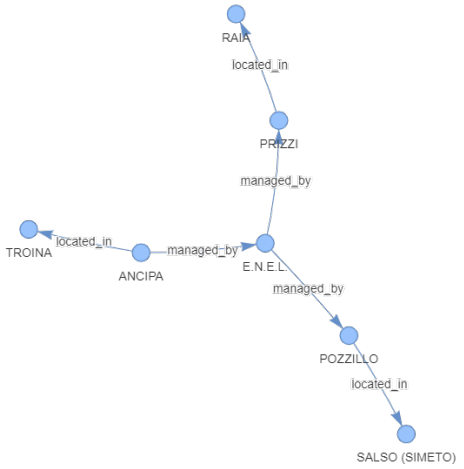

Fig. 1. Extracted knowledge graph

and a system prompt like this:

```
Extract entities from the following text
and return the result in JSON format.
Text: "{text}"
Expected format:
{{
    "Entity": ["Luca", "Marco"],
    "Relations": [
        {{"subject": "Luca",
          "object": "Marco",
          "relation": "friends"}}
    ]
}}
```

the LLM returns a list of entities and relationships, based on dam names, river names and managing organizations, with relations such as "located in" and "managed by", as shown in Figure 1.

V. Conclusion and Further Work

Besides structured data coming from public and private sources, information available as natural language from web pages (for instance about temporary events or from blogs and social media), is valuable for the touristic sector. In the upcoming phase of the project we are focusing on extracting well-formed NGSI-LD entities out of descriptive text in natural language, scraped from web sites or social media information sources that have touristic relevance.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Linking the World's Information*, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:10003259

[2] "NGSI-LD wikipedia," https://en.wikipedia.org/wiki/NGSI-LD, 2022, accessed: 2024-06-29.

[3] V. González, L. Martín, J. R. Santana, P. Sotres, J. Lanza, and L. Sánchez, "Reshaping smart cities through ngsi-ld enrichment," *Sensors*, vol. 24, no. 6, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/6/1858