Analisi dei Punti di Interesse nelle Isole

Pizzuto Salvatore Leonardo matr. 0696272
Messina Andrea matr. 0698481

Settembre 2022

Contents

1 Informazioni sul progetto

Elaborato prodotto come relazione di progetto per l'esame del corso "Tecniche per la gestione degli Open Data" tenuto dal prof. Davide Taibi presso l'Università degli studi di Palermo, indirizzo di Informatica.

2 Descrizione generale del progetto

Il progetto è stato sviluppato con l'intento di creare dei linked open data riguardanti i punti di interesse turistici (musei, aree archeologiche, monumenti) nelle due isole italiane principali, partendo da due dataset forniti dalle regioni stesse.

3 Dati di partenza e licenze

I dati sono forniti dalle regioni con la licenza (CC-BY 4.0)

- Dataset Sardegna:
  http://dati.regione.sardegna.it/dataset/anagrafe-degli-istituti-e-luoghi-della-cultura
- Dataset Sicilia:
  https://dati.regione.sicilia.it/dataset/musei-gallerie-e-siti-archeologici

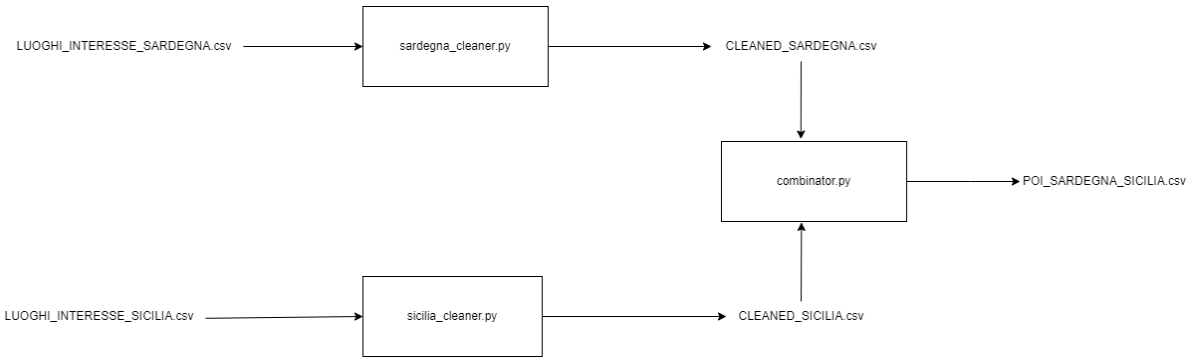
I dati forniti dal Ministero della Cultura (con cui successivamente verrà effettuato l'interlinking) sono rilasciati con licenza CC-BY-SA 4.0.
EndPoint Sparql Beni Culturali:
https://dati.cultura.gov.it/sparql

4 Note di rilascio dati

I dati prodotti dal progetto sono da considerarsi sotto licenza CC-BY-SA 4.0. Durante lo svolgimento del progetto, inoltre, è stata prodotta la mappa dei punti di interesse analizzati tramite uMap, rilasciata sotto licenza Open Data Commons Open Database License (ODbL). DataWrapper, invece, permette l'uso dei grafici solo dopo averli pubblicati sulla piattaforma.

5 Data cleaning

Si riporta di seguito uno schema indicativo che rappresenta il funzionamento della pipeline di elaborazione per tutto ciò che riguarda la pulizia dei dati:



Inseriamo inoltre tre sottosezioni nelle quali descriviamo il funzionamento dei programmi e delle loro funzioni.
Per ulteriori approfondimenti, vedere il codice opportunamente commentato.

5.1 sardegna_cleaner.py

Programma che si occupa di sgrezzare il dataset della Sardegna.
Di seguito elenchiamo le funzioni utilizzate:

- create_cleaned_sardegna_data():
  funzione principale che si occupa di coordinare il cleaning del dataset.
- delete_incomplete_data_sardegna(): funzione per eliminare i POI con valore nullo in "Prezzo" (FRBI).
- filter_open_places_sardegna(): funzione utile per il filtraggio dei POI aperti.
- fix_columns_sardegna(): eliminazione colonne inutili, gestione dei valori nulli, rinominazione colonne.
- cast_prices_sardegna(): funzione per conversione a float dei prezzi.
- fix_fax(): sistemazione formato numeri di telefono e fax.

5.2 sicilia_cleaner.py

Programma che si occupa di sgrezzare il dataset della Sicilia.
Di seguito elenchiamo le funzioni utilizzate:

- create_cleaned_sicilia_data():
  funzione principale che si occupa di coordinare il cleaning del dataset.
- filter_open_places_sicilia(): filtraggio dei POI aperti.
- fix_columns_sicilia(): pulizia delle colonne non utili .
- delete_incomplete_data_sicilia(): rimozione dei POI con un valore nullo in "Prezzo".
- fix_prices_sicilia(): sistemazione dei prezzi nel dataset (es. Sostituisci Gratuito/Ingresso Libero con 0.0) e casting a float dei prezzi.
- fix_cities_sicilia(): Sistemazione riga per riga di alcune denominazioni dei comuni.
- fix_categories_sicilia(): Raggruppamento delle categorie del dataset Sicilia secondo le tre categorie individuate (Musei,Gallerie,Raccolte , Aree archeologiche...).
- fix_addresses_sicilia(): sistemazione e uniformizzazione degli indirizzi, propedeutico alla funzione retrieve_lat_long_from_addresses_sicilia().
- retrieve_lat_long_from_addresses_sicilia(): estrazione delle coordinate geografiche a partire dall'indirizzo tramite geoPy.
- fix_phone_numbers_sicilia(): sistemazione formato numeri di telefono e gestione dei numeri non registrati.

5.3 combinator.py

Programma che si occupa di legare i due dataset precedentemente puliti. L'unica funzione utilizzata è combine_data(), che concatena i due dataset e memorizza il risultato in un nuovo file quale POI_SARDEGNA_SICILIA.csv .

6 Ontologia

Inizialmente era stata sviluppata tramite il software Protégé un'ontologia OWL (Web Ontology Language) sulla base dei dati trattati. Successivamente, è stato ritenuto opportuno utilizzare un'ontologia già esistente che modellasse in maniera più solida il contesto dei punti di interesse turistici.

L'ontologia utilizzata si può trovare al seguente link
https://dati.cultura.gov.it/cultural-ON/ENG.html

Lo schema mostrato dalla pagina, tuttavia, risulta incompleto. Per comprendere appieno ogni aspetto dell'ontologia è necessario leggere approfonditamente tutta la documentazione.

Successivamente, per controllo, sono state effettuate delle query di prova allo Sparql Endpoint del sito del Ministero della Cultura, del quale viene allegato il link nel paragrafo 3.
Ontologie utilizzate in combinazione con CULTURAL-ON

- CLVAPIT
  https://ontopia-lodview.agid.gov.it/onto/CLV
- GEO
  http://www.w3.org/2003/01/geo/wgs84_pos
- OWL
  http://www.w3.org/2002/07/owl

7 Struttura delle URIs

E' stato usato purl.org per generare delle URI persistenti. A livello di codice sono state utilizzate le funzioni urify_string() e our_urify_string() per sostituire i caratteri speciali con "_", inoltre viene utilizzato il package unidecode per sostituire i caratteri accentati coi rispettivi non accentati. La struttura delle URI utilizzata viene di seguito riportata:

- base_domain: https://purl.archive.org/purl/net/poi_sardegna_sicilia/risorse
- region_uri: base_domain/regioni/*Nome_Regione*
- city_uri: base_domain/comuni/*Nome_Comune*
- address_uri: base_domain/indirizzi/*Indirizzo*
- site_uri: base_domain/sedi/*Nome_Sede_Contatti_Indirizzo_SITE*
- cultural_institute_uri: base_domain/sedi/*Nome_Sede_Contatti_Indirizzo_INSTITUTE*
- contact_point_uri: base_domain/punti_contattabili/*Nome_Sede_Contatti_Indirizzo_CONTACTS*
- price_specification_uri: base_domain/specifiche_prezzo/*Nome_Sede_Contatti_Indirizzo_PRICESPEC*
- offer_uri: base_domain/offerte/*Nome_Sede_Contatti_Indirizzo_OFFER*
- ticket_uri: base_domain/biglietti/*Nome_Sede_Contatti_Indirizzo_TICKET*
- URI_TELEFONO (usate per hasTelephone): base_domain/contatti/*Contatto*

8 Creazione RDF e Interlinking

Una volta definita l'ontologia, la pipeline di elaborazione si conclude creando il file RDF/Turtle. Per la creazione del file TURTLE è stato usato il programma RDF_triples_generator_CULTURAL_ON.py, il quale tratta i dati del file POI_SARDEGNA_SICILIA.csv per produrre il nostro grafo di conoscenza, tramite la libreria python rdflib. Il file risultante sarà POI_RDF_TURTLE.ttl.

All'interno di questo file i dati precedentemente estratti sono collegati con i dati di DbPedia e con il catalogo dati del Ministero della Cultura.
Per verificare la presenza di Regioni e Comuni all'interno di DbPedia, è stata usata la seguente query sparql:

```
def is_city(urificated_city):
    uri = URIRef('http://dbpedia.org/resource/' + urificated_city)
    pp = URIRef('http://dbpedia.org/ontology/PopulatedPlace')
    g_temp = Graph()
    g_temp.parse(uri)
    try:
        response = g_temp.query(
            "ASK {?uri a ?pp}",
            initBindings={'uri': uri, 'pp': pp}
        )
    except HTTPError:
        return False
    print(str(uri) + " is a PopulatedPlace? " + str(response.askAnswer))

    return response.askAnswer
```

Le uniche due eccezioni sono state per i comuni di Siracusa (presente su DbPedia come Syracuse,_Sicily) e Campobello di Mazara (Non risulta un PopulatedPlace per DbPedia, proprietà usata per classificare e cercare i comuni).

Per effettuare l'interlinking con il catalogo dati del Ministero della Cultura è stata utilizzata una query sparql per estrarre i nomi istituzionali e le loro URI in modo che in caso di matching per nome istituzionale venisse indicata la URI recuperata come URI per l'interlinking.

```
def retrieve_institutes():
    sparql = SPARQLWrapper("https://dati.beniculturali.it/sparql") #Querying a remote SPARQL endpoint
    sparql.setQuery("""
        SELECT DISTINCT ?i ?l
        WHERE {
            ?i rdf:type cis:CulturalInstituteOrSite ;
               rdfs:label ?l ;
               cis:hasSite ?s .
            ?s cis:siteAddress ?a .
            ?a clvapit:hasRegion ?r .
            ?r rdfs:label ?region .
            FILTER(?region = "Sicilia" || ?region = "Sardegna")
        }
    """)

    sparql.setReturnFormat(JSON)
    try:
        result = sparql.query().convert()
    except HTTPError:
        return False

    return result
```

L'ultima query è stata eseguita utilizzando SPARQLWrapper, interfaccia per python utile a interrogare uno SPARQL Endpoint remoto.

9 Statistiche

9.1 Query SPARQL

Grazie a delle query SPARQL, siamo stati in grado di estrarre diversi dati interessanti riguardo i punti di interesse nelle isole. Abbiamo quindi interrogato il nostro grafo di conoscenza in questo modo:

- Query che estrae i prezzi medi per tipo di POI raggruppando in base alla regione e al tipo di POI stesso.

```
def query1(graph):
    results = graph.query("""
        PREFIX cis: <http://dati.beniculturali.it/cis/>
        PREFIX clvapit: <https://ontopia-lodview.agid.gov.it/onto/CLV/>
        PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
        PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

        SELECT ?l ?type (((ROUND(AVG(?p) * 100)) / 100) AS ?average)
        WHERE {
            ?o rdf:type cis:Offer ;
               cis:hasPriceSpecification ?p_spec;
               cis:includes ?t .
            ?p_spec cis:hasCurrencyValue ?p .
            ?t cis:forAccessTo ?c .
            ?c cis:hasSite ?s ;
               cis:hasCISType ?type .
            ?s cis:hasAddress ?a .
            ?a clvapit:hasRegion ?r .
            ?r rdfs:label ?l
        }GROUP BY ?r ?type
        ORDER BY ?l ?type
    """)

    return results
```

- Query che estrae il numero totale di POI per tipo raggruppando in base alla regione e al tipo di POI stesso.

```
def query2(graph):
    results = graph.query("""
        PREFIX cis: <http://dati.beniculturali.it/cis/>
        PREFIX clvapit: <https://ontopia-lodview.agid.gov.it/onto/CLV/>
        PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
        PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

        SELECT ?l ?type (COUNT(?c) AS ?tot)
        WHERE {
            ?c rdf:type cis:CulturalInstituteOrSite ;
               cis:hasSite ?s ;
               cis:hasCISType ?type .
            ?s cis:hasAddress ?a .
            ?a clvapit:hasRegion ?r .
            ?r rdfs:label ?l
        }GROUP BY ?r ?type
        ORDER BY ?l ?type
    """)

    return results
```

Con i dati estrapolati dalle query è stato quindi creato il file POI_STATS.csv, il quale è stato poi utilizzato come base per la creazione dei grafici tramite DataWrapper

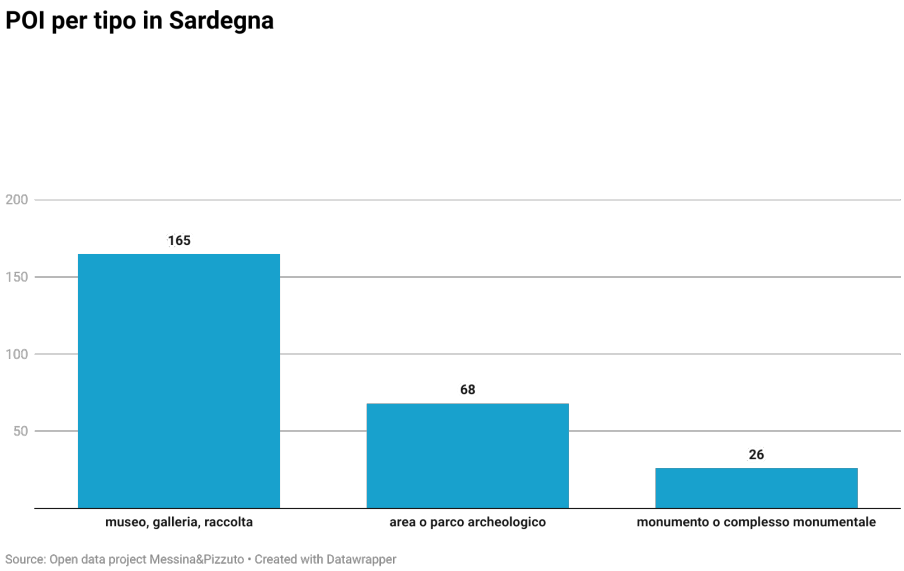
9.2 Statistiche elaborate e Grafici

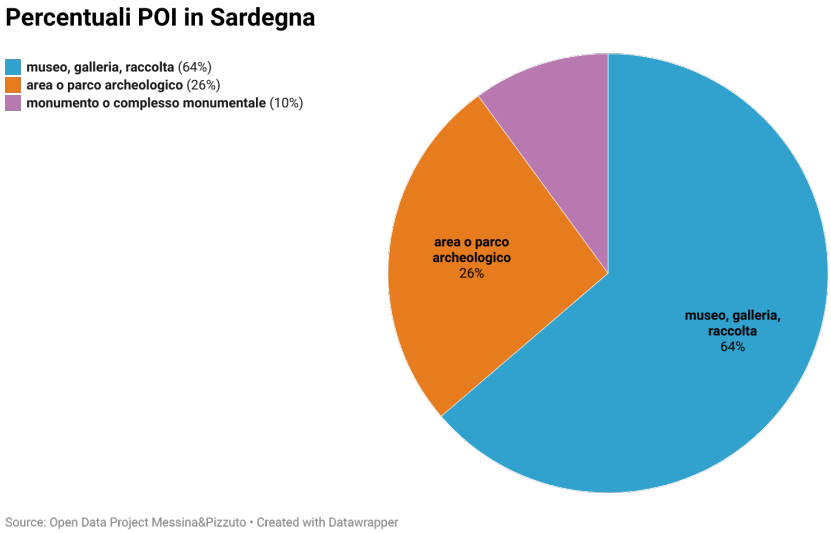
I dati estrapolati escludono tutte quelle voci nei dataset di partenza in cui i prezzi risultavano nulli o comunque non indicati, per poter effettuare un'analisi successiva sui prezzi. Ciò significa che queste statistiche sono influenzate anche dall'assenza di determinate voci all'interno del dataset finale.

I primi grafici prodotti riguardano il totale dei POI, divisi per tipo, in Sardegna e Sicilia

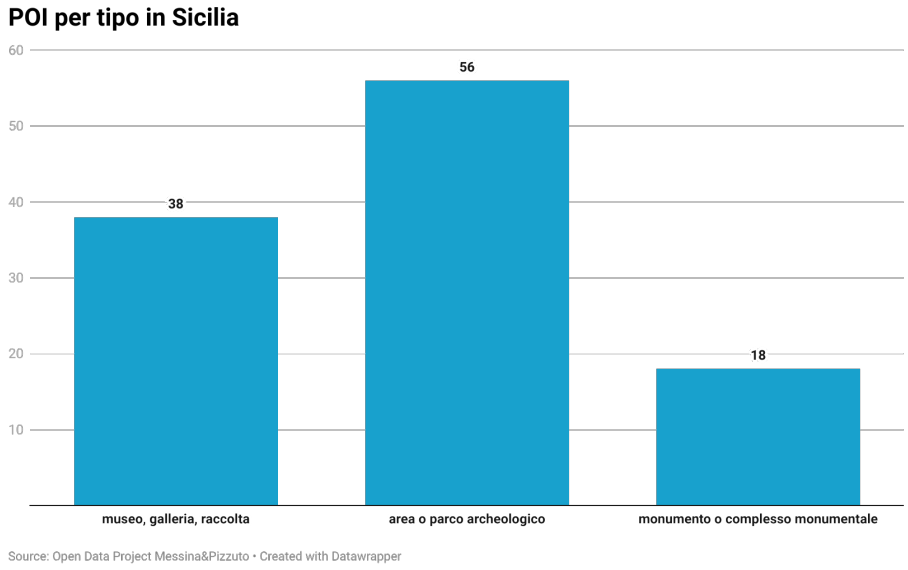
9.2.1 Analisi POI per regione

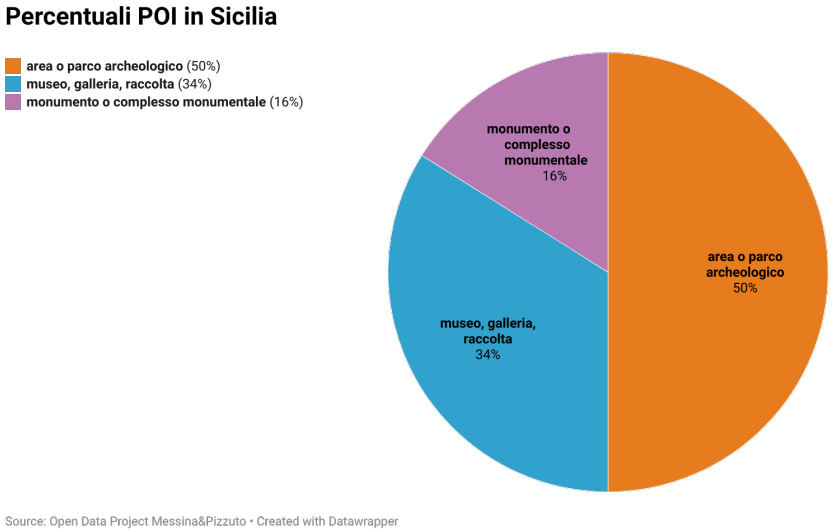
La prima regione da analizzare osservandone i grafici è la Sardegna:





Da questi grafici viene evidenziato che i musei/gallerie/raccolte rappresentano più del 50% (165 unità sul totale) dei punti di interesse in Sardegna, mentre i monumenti solamente il 10% (26 unità sul totale).

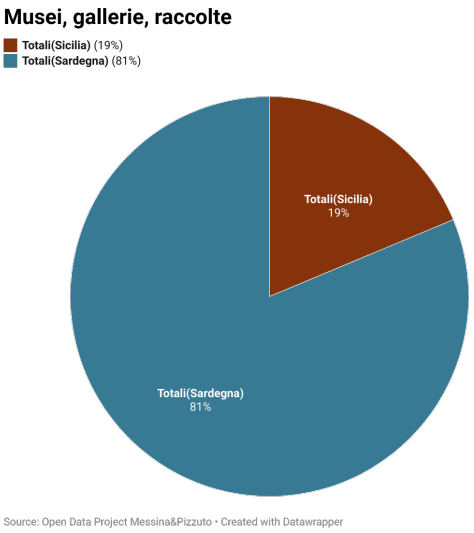




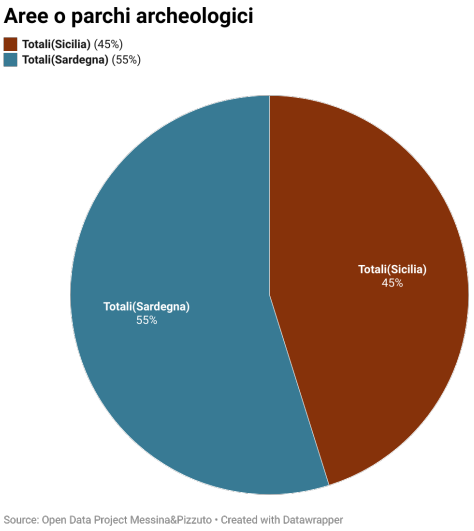
Questi grafici mostrano, invece, che i parchi archeologici rappresentano il 50% (56 unità sul totale) dei punti di interesse in Sicilia, mentre i monumenti solamente il 16% (18 unità sul totale).

9.2.2 Confronto per tipo di POI fra le due regioni

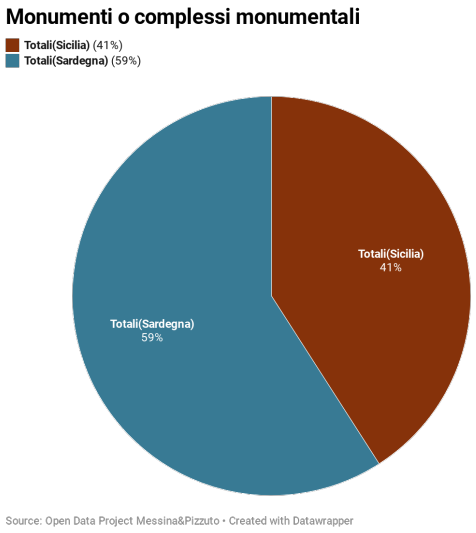
Iniziamo l'analisi parlando del grafico di Musei, Gallerie e Raccolte, il quale mostra un distacco netto in favore della Sardegna



Successivamente osserviamo il confronto tra regioni per POI di tipo Aree o Parchi Archeologici, il quale mostra una differenza relativamente sottile fra Sardegna e Sicilia (55% vs 45%)



Ed infine i Monumenti o Complessi Monumentali, nei quali troviamo sempre una differenza non esageratamente grande fra le due regioni, con un vantaggio nei confronti della Sardegna.



9.2.3 Aspetti Econonomici

Per concludere l'analisi delle statistiche elaborate, presentiamo una tabella che mostra i prezzi medi per regione e tipo di POI.

Prezzi medi Sardegna-Sicilia per tipologia di POI

POI	museo, galleria, raccolta	area o parco archeologico	monumento o complesso monumentale
Prezzo medio in euro(Sardegna)	4.26	4.18	4.88
Prezzo medio in euro(Sicilia)	3.58	2.39	1.83

Source: Open Data Project Messina&Pizzuto • Created with Datawrapper

Ciò che si evince dalla tabella è che, mentre la differenza dei prezzi medi nei musei sono equiparabili fra le due regioni, nelle aree archeologiche e nei monumenti la differenza di prezzo è molto più alta (più di 1.50 euro), premiando così la Sicilia come regione più economica.

10 Altre Query SPARQL

Abbiamo scritto, inoltre, delle altre query per interrogare il nostro grafo di conoscenza RDF, di seguito riportate.

- Estrazione dei luoghi della cultura presenti nella città di Palermo.

```
def query3(graph):
    results = graph.query("""
            PREFIX cis: <http://dati.beniculturali.it/cis/>
            PREFIX clvapit: <https://ontopia-lodview.agid.gov.it/onto/CLV/>
            PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
            PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

            SELECT ?n
            WHERE {
                ?i rdf:type cis:CulturalInstituteOrSite ;
                   cis:hasSite ?s ;
                   cis:institutionalName ?n .
                ?s cis:hasAddress ?a .
                ?a clvapit:hasCity ?c .
                ?c rdfs:label ?l .
                FILTER (?l = "Palermo") .
            }
        """)

    for row in results:
        print(row)
```

- Estrazione dei luoghi della cultura che contengono nel nome istituzionale la parola "casa" (es. Casa Museo di Giovanni Verga).

```
def query4(graph):
    results = graph.query("""
                PREFIX cis: <http://dati.beniculturali.it/cis/>
                PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

                SELECT ?n
                WHERE {
                    ?i rdf:type cis:CulturalInstituteOrSite ;
                       cis:institutionalName ?n .
                    FILTER(REGEX(?n, "casa ", "i"))
                }
            """)

    for row in results:
        print(row)
```

- Estrazione dei nomi istituzionali e dei relativi numeri di telefono dei punti di interesse situati a Cagliari.

```
def query5(graph):
    results = graph.query("""
                PREFIX cis: <http://dati.beniculturali.it/cis/>
                PREFIX clvapit: <https://ontopia-lodview.agid.gov.it/onto/CLV/>
                PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

                SELECT ?n ?t
                WHERE {
                    ?i rdf:type cis:CulturalInstituteOrSite ;
                       cis:institutionalName ?n ;
                       cis:hasSite ?s ;
                       cis:hasContactPoint ?p .
                    ?p cis:hasTelephone ?t .
                    ?s cis:hasAddress ?a .
                    ?a clvapit:hasCity ?c .
                    ?c rdfs:label ?city .
                    FILTER(?city = "Cagliari")
                }
            """)

    for row in results:
        print(row)
```

- Estrazione delle denominazioni dei musei presenti in Sardegna

```
def query6(graph):
    results = graph.query("""
                    PREFIX cis: <http://dati.beniculturali.it/cis/>
                    PREFIX clvapit: <https://ontopia-lodview.agid.gov.it/onto/CLV/>
                    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
                    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

                    SELECT ?n
                    WHERE {
                       ?c rdf:type cis:CulturalInstituteOrSite ;
                          cis:institutionalName ?n ;
                          cis:hasCISType cis:Museum ;
                          cis:hasSite ?s .
                       ?s cis:hasAddress ?a .
                       ?a clvapit:hasRegion ?r .
                       ?r rdfs:label ?l .
                       FILTER (?l = "Sardegna")
                    }
                """)
    for row in results:
        print(row)
```

11 Mappa consultabile dei punti di interesse

Attraverso il sito uMap abbiamo generato una mappa dei punti di interesse sulla base dei dati trattati. La mappa è consultabile seguendo questo link: http://umap.openstreetmap.fr/it/map/poi_sardegna_sicilia_8041847/38.937/13.223. Ricordiamo che la mappa è rilasciata sotto Licenza ODbL. OpenStreetMap-Copyright.