KDI : **Knowledge and Data Integration**

‘COVID-19 Data Integration’

KDI Final Presentation

Contributors

- Data Scientist: Nisha Antony
- Data Scientist: Maria Jyate
- Domain Expert: Lorenzo Donini
- Knowledge Engineer: Lorenzo Donini
- Knowledge Engineer: Daniel Gotca
- Project Manager: Maria Jyate
- Tutor Data: Simone Bocca
- Tutor Knowledge: Mauro Dragoni

Table of Contents

1 **Project description**

2 **Resources**

3 **Problems and Solutions**

4 **Outcomes**

5 **Open Issues Future works**

Table of Contents

1. **Project description**
2. Resources
3. Problems and Solutions
4. Outcomes
5. Open Issues Future works

Project description

The goal of this group activity is to group all the implicit and explicit data about the global pandemic we are currently living in: Covid-19. Thanks to this integration, we aim to understand the diffusion of the virus in the Trentino Region, since localizing new epidemic centers of the virus is a fundamental factor to limit its diffusion.

In order to obtain a complete data collection, we integrated data about Trentino facilities and point of interest as well as data about the situation in the neighbor countries due to the tourists attraction present in the province.

Project description

The high number of tourist attraction lead to an higher risk of new infections: this means that it is also important to know how the pandemic situation is in those countries. It is relevant to understand not only the situation, but how other nations act in regard of Covid-19 virus. In a region that thrives on tourism the economic losses would be high if the virus is underestimated.

Table of Contents

1. Project description

2. **Resources**

3. Problems and Solutions

4. Outcomes

5. Open Issues Future works

Data resources

The datasets used were all in .csv and .xls format and proposed in a tabular way.

- COVID-19 Coronavirus data
- COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections
  - Reference_hospitalization_all_locs dataset
  - Summary_stats_all_locs.csv
- COVID-19 emergency health situation: Province of Trentino

Metadata

- In the Informal Modeling phase, we created a metadata table to describe the datasets and the attributes before data cleansing.
- The final metadata was created based on the DCAT standards.
- This metadata describes the filtered and formatted datasets which was the output of the Formal Modeling phase.
- The metadata is saved in a .JSON format with the datasets and attributes level descriptions.

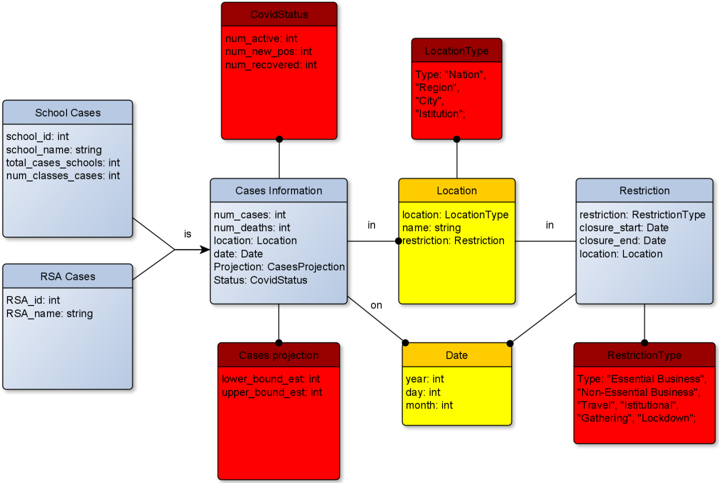
Knowledge resources

- **Ranjith** is a young man of 21 years old from India.
  Give the progress of COVID-19 in Trento & return last week data in the Province
- **Franco** is a middle-aged man who owns a nice hotel near ski slopes.
  Give the number of possible tourist in Trentino & return the prediction of cases and mobility in Trentino
- **Gianni** is a school principle of a complex of an elementary and a middle school.
  Give the number of COVID-19 cases in Trentino schools & return the number of school cases in the region

Knowledge resources

- **Marta** is an employee at the local transportation agency and her job is to manage and guarantee the safety and security of all the staff and passengers.
  Give the number of the cases in the neighbour countries & return the data about the last three days in those countries
- **Carla** is the manager at a private RSA.
  Give the number of cases in the Trentino RSAs & return the number of cases in the RSAs

Knowledge resources



CovidStatus
num_active: int
num_new_pos: int
num_recovered: int

LocationType
Type: "Nation",
"Region",
"City",
"Istitution";

School Cases
school_id: int
school_name: string
total_cases_schools: int
num_classes_cases: int

RSA Cases
RSA_id: int
RSA_name: string

Cases Information
num_cases: int
num_deaths: int
location: Location
date: Date
Projection: CasesProjection
Status: CovidStatus

Location
location: LocationType
name: string
restriction: Restriction

Restriction
restriction: RestrictionType
closure_start: Date
closure_end: Date
location: Location

Cases projection
lower_bound_est: int
upper_bound_est: int

Date
year: int
day: int
month: int

RestrictionType
Type: "Essential Business",
"Non-Essential Business",
"Travel", "Istitutional",
"Gathering", "Lockdown";

Figure: EER

Table of Contents

1 Project description

2 Resources

3 **Problems and Solutions**

4 Outcomes

5 Open Issues Future works

Problems and Solutions

- **Ontology: general incompatibility between KOS and Protégé.**
  Extraction of the RDF file from the KOS and manually insertion of the Etypes, the DataProperties, and ObjectProperties.
- **As stated in the this tutorial, the Common entities has to be inserted as individuals but this procedure was not supported by the KOS.**
  the Enumeration Types were defined as a class while the Individuals were defined as the DataProperties. This is the case of LocationType and RestrictionType.

Problems and Solutions

- **The Karmalinker tool wan not able to handle big datasets (more than 10.000 rows)**
  The big datasets were partitioned in smaller chunks and then loaded in the Karmalinker workspace, to create the mappings and the EML and RDF files.

Table of Contents

1. Project description

2. Resources

3. Problems and Solutions

4. **Outcomes**

5. Open Issues Future works

Outcomes

- The project developed following the proposed methodology can be considered successful in its entirety since the resulting knowledge graph.
- The KG produced is well populated and contains all the information needed in order to answer to the CQs initially defined for our personas.
- The ontology can be further expanded with more specific information about the COVID-19 such as, for example, diagnoses and symptomatic information.

Outcomes

CLASS OF THE REFERENCE ONTOLOGY(α)	24
CLASS OF OUR ONTOLOGY (β)	10
COMMON CLASS (B)	6

Table: Number of classes of our and the ontology took in consideration [1]

Coverage	25,00 %
Flexibility	16,60%
Extensiveness	11,76%
Sparsity	53,00%

Table: Evaluation results

Table of Contents

1 Project description

2 Resources

3 Problems and Solutions

4 Outcomes

5 **Open Issues Future works**

Open Issues

Diversity

Since the COVID-19 is a relatively new fact, most of the datasets retrieved contained similar information. Almost all of the datasets that we collected contains similar information in different format making things difficult for us to find answers for the CQs.

KDI Knowledge and Data Integration

‘COVID-19 Data Integration’

KDI Final Presentation