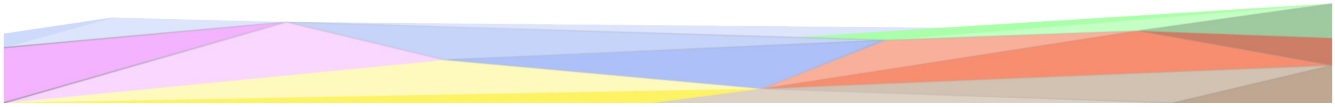
Dipartimento di Ingegneria e Scienza dell'Informazione

– KnowDive Group –

Integration of medical data on Covid-19

Document Data:	Reference Persons:
October 21, 2020	Nisha Antony
	Daniel Gotca
	Maria Jyate
	Lorenzo Donini



Contents

1 Knowledge Graph Codebook **1**

1.1 Knowledge Graph general description 1

1.2 Data level 1

1.2.1 Datasets general details 1

1.2.2 Datasets metadata documentation 3

1.3 Ontology level 4

1.3.1 Ontology general details 4

1.3.2 Ontology metadata documentation 4

1.4 Knowledge Graph Evaluation 7

1 Knowledge Graph Codebook

The first of the two sections, in the current document, contains the codebook of the whole KG (Knowledge Graph), including the description of all the data and information that it contains.

1.1 Knowledge Graph general description

This sub section aims to give a general description of the KG, reporting:

- the context/domain in which the KG lives and works;
- *The Problem* the KG aims to solve;
- How the KG can solve *The Problem*

This year world population has been afflicted by the Covid-19 global pandemic, everyday the situation changes, the number of active cases, recovered and deaths changes and this generates tons of data which can be integrated and studied for many purposes, starting from the progress of the epidemic to the measures given by the governments to stop the spreading. The problem nowadays is that there are multiple databases with different data, from small local areas to bigger places which are disconnected one from another, so there is no possibility to have multiple information regarding your needs without retrieving information from different places. The goal of our KG is to integrate the different data starting from the single places here in Trentino like RSA or Schools to the whole country and the nearby ones in order to give complete information to a wide range of people about the situation, the restrictions in the requested period and also predictions.

1.2 Data level

The data level section aims to describe in details the (final version of) datasets collected and managed by the KG, with a description of each variable involved.

1.2.1 Datasets general details

The datasets, collected in .csv and .xls format, were provided by various health and research organizations. The collected data was produced as an outcome of the research and observation made by these organisations during the first wave of the pandemic and it covers almost all the aspect of Covid-19, from cases information to restrictions, and hospital admissions and so on. The Datasets used in this project are listed below:

- COVID-19 Coronavirus data [3]

Field name	Description
Title	Covid-19 Coronavirus data
Description	The European Centre for Disease Prevention and Control(ECDC) has created this dataset by collecting reports from health authorities worldwide ever since the Covid-19 outbreak. The data is updated on a daily basis and is available in CSV format. The dataset contains cumulative daily status regarding the pandemic across the world.

Category	Health
Keywords	COVID-19, disease outbreak, corona virus, SARS-CoV-2, coronavirus, severe acute respiratory syndrome coronavirus-2
Last update	2020-10-20
Publisher	European Centre for Disease Prevention and Control (ECDC)
Contact Information	ECDC European Centre for Disease Prevention and Control https://www.ecdc.europa.eu/en/about-ecdc
Frequency	Daily
Temporal Coverage	2019-12-31
Spatial Coverage	Europe, Asia, Africa, America, Oceania

- COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections [4]

Field Name	Description
Title	COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections
Description	IHME's COVID-19 projections were developed in response to requests from the University of Washington School of Medicine and other US hospital systems and state governments working to determine when COVID-19 would overwhelm their ability to care for patients. The forecasts show demand for hospital services, daily and cumulative deaths due to COVID-19, rates of infection and testing, and the impact of social distancing, organized by country and state (for select locations).
Category	Health
Keywords	COVID-19, disease outbreak, coronavirus
Last update	2020-10-15
Publisher	Institute for Health Metrics and Evaluation (IHME).
Contact Information	covid19@healthdata.org
Temporal Coverage	2020-02-04
Spatial Coverage	Global

- COVID-19 emergency health situation: Province of Trentino [1]

Field Name	Description
Title	COVID-19 emergency health situation: Province of Trentino
Description	The datasets include the clinical status and current situation of Trentino municipalities. The data is in Italian and is available in CSV format. The information is collected by the Trentino Digitale from the health centres to keep residents up-to-date regarding the Covid situation.

Category	Health
Keywords	COVID-19, coronavirus, Trentino data, Covid-19 Trentino
Last update	2020-10-20
Publisher	Trentino Digitale
Contact Information	https://www.trentinodigitale.it/
Spatial Coverage	Province of Trentino

1.2.2 Datasets metadata documentation

In this section are reported the metadata at dataset attribute level, through a description of each variable involved in the datasets collected, specifying the variable types, meanings, value-set (possible values), and every other meaningful variable information.

- COVID-data-integration-1.in.json

Attributes after filtering	Description	Type	Data definition
cases	Number of new cases on the date of the report	int	Core
deaths	Number of deaths on the date of the report	int	Core
countriesAndTerritories	Country and territory name related to the records	string	Contextual
year	Year from the date of report	int	Common
month	Month from the date of report	int	Common
day	Day from the date of report	int	Common

Table 4: COVID-19 Coronavirus data metadata after filtering

- COVID-19 Mortality, Infection, Testing, Hospital Resource Use, and Social Distancing Projections
  - Reference_hospitalization_all_locs dataset

Attributes after filtering	Description	Type	Data definition
location_name	Name of the country or subnational location	string	Common
est_infections_lower	Lower uncertainty bound of estimated infections	int	Contextual
est_infections_upper	Upper uncertainty bound of estimated infections	int	Contextual
year	Year from the date of report	int	Common
month	Month from the date of report	int	Common
day	Day from the date of report	int	Common

Table 5: Metadata description of Reference_hospitalization_all_locs dataset

  - Summary_stats_all_locs.csv

Attributes after filtering	Description	Type	Data definition
location_name	Name of the country or subnational location	string	Common
restriction_type	Type of restriction	string	Contextual
closure_start	Date of the closure start	date	Core
closure_end	Date of the closure end	date	Core

Table 6: Metadata description for Summary_stats_all_locs.csv

- COVID-19 emergency health situation: Province of Trentino

Attributes after filtering	Description	Type	Data definition
guariti	Number of recovered cases	int	Contextual
deceduti	Number of deaths	int	Core
pos_att	Number of active cases	int	Contextual
rsa	Number of cases in rsa	int	Core
year	Year from the date of report	int	Common
month	Month from the date of report	int	Common
day	Day from the date of report	int	Common

Table 7: Metadata description for Stato_clinica_td.csv

1.3 Ontology level

The ontology level section aims to describe the underlying KG ontology, through the description of its elements at each level, reporting so the language, conceptual and schema resources used within it.

1.3.1 Ontology general details

Starting from the EER model, we want to obtain the Schema Knowledge Graph (SKG Figure **??**), which is based on a data driven data-driven requirement and uses standards already implemented thanks to the teleologies.
The graph-like structure allow us to clearly see how the data interact with each other, with the nodes representing the entities and their attributes while the relations between these elements are viewed as edges. Both the entities and the relations are constrained by properties and logics already defined at the EER schema level.
This section aims to illustrate how the ontology was created. In order to do that tools such as Protege have been used to deliver the final result (more on that in the section below) . Starting from the EER model, which give a schematic representation of the structure, we represented the same data in a more complete and detailed way by building the ontology to entirely represent the database.
We used Protégé [6] to add and modify the necessary classes with the corresponding concepts.Then we added the required data and object properties regarding the Etypes present with the procedure already mentioned above. To add the enumeration we adopted a different method from the one published in this tutorial [5] since that methodology was not supported by the KOS system: so a different methodology has been used, the Enumeration Types were defined as a class while the Individuals were defined as the class attributes.

1.3.2 Ontology metadata documentation

In this section instead, are reported the more specific metadata describing the single elements of the ontology (terms, concepts, ETypes and relations).

	Present	Added by us	Total
Classes	3	7	10
DataProperties	6	17	23
ObjectProperties	0	10	10

KOS CONCEPT	KOS CONCEPT ID	COMMENT
School_Cases	120042	representation of covid-19 cases in school
RSA_Cases	120043	representation of covid-19 cases in residence healthcare assistant
Cases_Information	120044	representation of covid-19 information in general
CovidStatus	120045	representation of the actual covid-19 situation
CasesProjection	120046	representation of the projection of cases of covid-19 in the future
LocationType	120047	specification of the location investigate
RestrictionType	120048	specification of the type of restriction
Location	132	a point or extent in space
Date	103420	assign a date to
Restriction	31867	a principle that limits the extent of something

Table 8: Classes present in the ontology

CONCEPT NAME	CONCEPT GID	CLASS	CLASS ID
has_day	81001	Date	103420
has_year	80976	Date	103420
has_elevation	28270	Location	132
has_latitude	46263	Location	132
has_longitude	46270	Location	132
has_period	80647	Restriction	31867
has_Essential_Business	120087	RestrictionType	120048
has_Gathering	43384	RestrictionType	120048
has_Istitutional	120089	RestrictionType	120048
has_Lockdown	5912	RestrictionType	120048
has_Non_Essential_Business	120088	RestrictionType	120048
has_Travel	109441	RestrictionType	120048
has_Region	46452	Locationtype	120047
has_Nation	44214	Locationtype	120047
has_Istitution	120090	Locationtype	120047
has_City	45988	Locationtype	120047
has_num_active	120091	Covidstatus	120045
has_num_new_pos	120092	Covidstatus	120045
has_num_recovered	120093	Covidstatus	120045
has_lower_bound_pos	120094	Casesprojection	120046
has_upper_bound_pos	120095	Casesprojection	120046
has_RSA_ID	120102	Rsa_cases	120043
has_RSA_name	120103	Rsa_cases	120043
has_school_id	120098	School_cases	120042
has_school_name	120099	School_cases	120042
has_total_cases_schools	120100	School_cases	120042
has_num_classes_cases	120101	School_cases	120042
has_num_deaths	120097	CasesInformation	120044
has_num_cases	120096	CasesInformation	120044
has_num_deaths	120097	CasesInformation	120044

Table 9: Data properties

CONCEPT NAME	DOMAIN	DOMAIN GID	RANGE	RANGE ID
has_has	CasesInformation	120044	Covidstatus	120045
has_has	CasesInformation	120044	Casesprojection	120046
has_has	Location	132	Locationtype	120047
has_has	Restriction	31867	RestrictionType	120048
has_is_in	CasesInformation	120044	Location	132
has_is_in	Restriction	31867	Location	132
has_is_on	CasesInformation	120044	Date	103420
has_is_on	Restriction	31867	Date	103420
has_is_RSACases	CasesInformation	120044	Rsa_cases	120043
has_is_SchoolCases	CasesInformation	120044	School_cases	120042

Table 10: Object properties

1.4 Knowledge Graph Evaluation

CLASS OF THE REFERENCE ONTOLOGY (α)[2]	24
CLASS OF OUR ONTOLOGY (β)	10
COMMON CLASS (B)	6

Table 11: Parameters [2]

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

Table 12: Evaluation metrics

References

[1] Trentino Digitale, *Covid-19 emergency health situation: Province of trentino.*

[2] B. Dutta and M. DeBellis, *Codo: Codo: an ontology for collection and analysis of covid-19 data.*, 2020, In Proc. of 12th Int. Conf. on Knowledge Engineering and Ontology Development (KEOD), 2-4 November 2020 (accepted).

[3] European Centre for Disease Prevention and Control (ECDC), *Covid-19 coronavirus data.*

[4] Institute for Health Metrics and Evaluation (IHME), *Covid-19 mortality, infection, testing, hospital resource use, and social distancing projections.*

[5] Matthew Horridge, Simon Jupp, Georgina Moulton, Alan Rector, Robert Stevens, and Chris Wroe, *A practical guide to building owl ontologies using protégé 4 and co-ode tools edition1. 2*, The university of Manchester **107** (2009).

[6] Stanford University, *Webprotégé.*